Making Sense of the Genetic Code with the Path-Distance Model
Based on RNA-dependent Pathways

Brian K. Davis

Research Foundation of Southern California,
8837 Villa La Jolla Drive, #13595, La Jolla,
CA 920309, U.S.A.

Summary. Free α-carboxyl distribution in amino acid biosynthesis, genetic code domains, and pre-divergence tRNA phylogenetics conserve the imprint of tRNA-dependent amino acid synthesis pathways during code formation. Their dicarboxyl distribution is linked here to tRNA cofactor exchange credited with anomalies apparent in $tRNA^{leucine}$ and $tRNA^{arginine}$ coding specificity. Pre-species-divergence tRNA specific for amino acids synthesized from pyruvate, phosphoenolpyruvate, and phospho-glycerate exhibit elevated identity with $tRNA^{asparagine}$, consistent with oxaloacetate being an upstream precursor before the protein takeover of these pathways. As reaction segments connecting extant precursors to oxaloacetate predated amino acid synthesis, path-distances in reconstructed tRNA-dependent pathways effectively matched those in biosynthesis. The path-distance principle of code-formation equating these distances with amino acid coding-order was shown to accommodate over fifty code features. RNA-based coding specificity, reliant on tRNA path-identity elements, led to an explanation for class duality in synthetase enzymes and codon 5'-base invariance among same-family amino acids. Path-distance evidence revealed the first proteins contained four NH_4^+ fixer/N-donor residues - aspartate, glutamate, asparagine, glutamine - assigned XAN triplets (X, coding site, N, degenerate site). Residue potentials revealed they could produce α-helices, β-turns, and proto-enzymes. β-Sheets and acid-base catalysis arose later, as codon mid- and 3'-base were successively recruited. Path-distance distributions revealed clusters of polar and non-polar residues formed at different stages of code formation.

1. Introduction

It became possible to unify, for the first time, more than twenty seemingly unrelated structural regularities attributed to the genetic code, when the number of reaction steps in amino acid biosynthesis was equated to the time-order of their entry into the genetic code [1, 2]. These regularities included conspicuous, non-overlapping clusters of polar and non-polar amino acids [3-6], 5'-base invariance among codons for same-family amino acids, and allocation of a codon 4-set (3'-base degenerate) to each of the six smallest amino acid residues in proteins [7]. Unlike most biosynthesis pathways, however, code structure and tRNA species, pre-dating the Last Common Ancestor (LCA), provide compelling evidence that amino acid synthesis pathways were tRNA-dependent during code formation [8].

Specifically, five domains and three small quasi-domains form the standard code, with each domain containing a distinct combination of same-family amino acids, structurally related tRNA, and nearest-neighbor codons [2, 8]. tRNA within the same domain share the same core structure group [8, 9] and the conserved imprint of their pre-LCA base sequences established they had diversified from a common ancestral tRNA [8]. Extensive utilization of dual cofactor/adaptor tRNA in amino acid synthesis during code formation [8, 10] clarified the source of the correlation observed between amino acid synthetic-order and coding-order [1, 2]. Furthermore, existence of path-specific elements within prokaryote tRNA cofactors in asparagine (Asn) and glutamine (Gln) synthesis [11] implies, in view of these developments, a general RNA-based coding mechanism for specifically matching amino acids with their base triplets, preceding protein synthetases [10].

tRNA-dependent amino acid synthesis pathways have been reconstructed in this study, using biosynthesis reaction sequences and pre-LCA tRNA phylogenetics. In addition to providing an amino acid synthetic-order contemporaneous with code formation, the reaction sequences obtained will be seen to resolve anomalies evident in the coding of leucine (Leu) and arginine (Arg). A new insight arose into the nature of the pathways of central metabolism that gave rise to the amino acid synthesis pathways. Code evolution also has been broadly re-interpreted in the context of tRNA-dependent amino acid synthesis. An annotated list of more than fifty features attributed to the code is attached (Table S1), illustrating the scope of the path-distance principle of code formation.

2. Scope of tRNA-dependent amino acid synthesis during code formation

RNA-dependent amino acid synthesis pathways occur in all three species domains (Fig. 1a). A tRNA cofactor participates in selenocysteine (Sec) synthesis in Archaea, Bacteria, and Eukarya, placing this pathway in the pre-LCA era. In some bacteria, cysteine (Cys) synthesis retains a Sep-tRNA^{Cys} intermediate (Sep, phosphoserine). Prokaryotes commonly synthesize asparagine (Asn) and glutamine (Gln) on a tRNA-dependent pathway. Use of tRNA cofactors in prokaryote synthesis of Asn^2 and Gln^2 (superscripts signify path-length) [1] is consistent with their being 'the protected root' [12] of a once extensive network of pre-LCA tRNA-dependent amino acid synthesis pathways. Takeover of nearly all amino acid synthesis pathways by proteins apparently extinguished the cofactor function of most tRNA species. The strongest evidence for tRNA participation in early amino acid synthesis is accordingly found conserved within the genetic code, pre-LCA tRNA base sequences, and, as shown in the following two sections, amino acid synthesis pathways.

Figure 1b portrays the genetic code domains. Each spans a region of contiguous codons read by tRNA, with the same core group [8, 9], specific for amino acids sharing the same precursor in central metabolism. Codon contiguity within code domains has a probability, $p = 2.16 \times 10^{-6}$, it occurred by chance [8]. The probability same-family amino

(a)

transition	B	A	E
$\text{Asp-tRNA}^{\text{Asn}} \rightarrow \text{Asn-tRNA}^{\text{Asn}}$	++	++	-
$\text{Glu-tRNA}^{\text{Gln}} \rightarrow \text{Gln-tRNA}^{\text{Gln}}$	++	+++	-
$\text{Sep-tRNA}^{\text{Cys}} \rightarrow \text{Cys-tRNA}^{\text{Cys}}$	+	-	-
$\text{Ser-tRNA}^{\text{Sec}} \rightarrow \text{Sec-tRNA}^{\text{Sec a}}$	+++	+++	+++

(b)

5'↓ mid →	U	C	A	G	↓ 3'
U	Phe IB	Ser II	Tyr IB	Cys IA'	U
					C
	Leu II		Ter	Ter	A
				Trp IA'	G
C	Leu II	Pro ID	His ID	Arg IA	U
					C
			Gln ID		A
					G
A	Ile IA	Thr IA	Asn IA	Ser II	U
					C
			Lys IA	Arg IA	A
	Met IA				G
G	Val IA'	Ala IA	Asp ID	Gly ID	U
					C
			Glu ID		A
					G

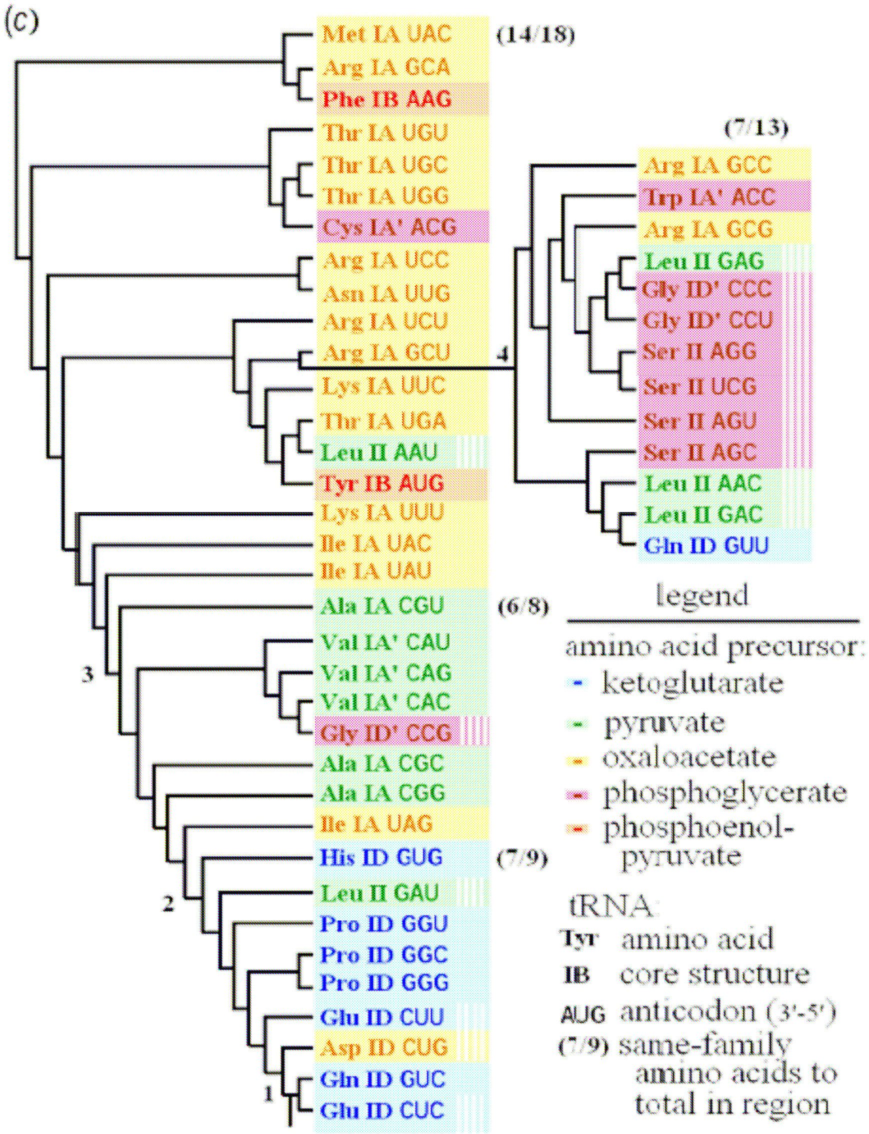


Figure 1. Evidence for pre-LCA tRNA-dependent amino acid synthesis. (*a*) Distribution of tRNA-dependent amino acid synthesis among Archaea (A), Bacteria (B), and Eukarya (E). Selenocysteine synthesis is tRNA-dependent in species from all three domains. Prokaryotes commonly utilize a tRNA cofactor in Asn and Gln synthesis. Cysteine synthesis has a phospho-seryl-tRNA intermediate, Sep-tRNA^{Cys}, in some bacteria. +++, all relevant species,; ++, most; +, some; and -, none. [a] Archaea and Eukarya species utilize Sep-tRNA^{Sec} [13]. (*b*) Domains within the standard code. Same-family amino acids, nearest-neighbor codons, and tRNA of the same type and subtype characterize each domain. Five domains and three small quasi-domains form the code. Solid colors designate a domain and stripes a quasi-domain. Yellow, amino acids from oxaloacetate in biosynthesis; blue, ketoglutarate; green, pyruvate; rose, 3-phosphoglycerate, and tan, phosphoenolpyruvate. Roman numerals and letters specify tRNA type and subtype [8, 9]. (*c*) Cladogram constructed from the conserved imprint of pre-LCA tRNA sequences, analyzed by the neighbour-joining method [8]. tRNA species for same-family amino acids cluster in the same tree region. Four tree regions are identified (numbered), with tRNA mainly from the same code domain. Colors designate, as in (b), code domain and amino acid precursor.

acids acquired tRNA with the same core group was only, $p = 8.45 \times 10^{-4}$ [8]. These regularities in code structure reveal synthetically related amino acids acquired structurally related tRNA, cognate with contiguous codons. Code structure and pre-LCA tRNA sequences, consequently, conserve compelling evidence that tRNA cofactors participated in amino acid synthesis during code formation [8]. tRNA diversification, new amino acid synthesis, and codon recruitment, furthermore, were coordinated.

As two-thirds of tRNA sequence variability arose in the long interval following species-divergence [14], ‘noise’ from this source was filtered-out before seeking to establish the phylogenetic relationship between pre-LCA tRNA paralogs [8]. A cladogram obtained on analysis of the conserved imprint of pre-LCA tRNA using the neighbor-joining method [8], at non-universal tRNA sites jointly conserved in the consensus sequence from sources in

Archaea, Bacteria, and Eukarya, contains four distinct regions of same-family amino acids (Fig. 1c). The tRNA tree root is placed among type-ID tRNA for NH_4^+ fixer/N donor amino acids Gln^2, glutamate (Glu^1), and aspartate (Asp^1). tRNA specificity for same-family amino acids within each region had a Kendall correlation (corrected for tied pairs) of $\tau = 0.92$, and $p = 7.0 \times 10^{-9}$ [8]. tRNA in tree regions and code domains correlated strongly; $\tau = 0.72$, with $p = 8.3 \times 10^{-6}$. Codon contiguity in tree regions and code domains was also highly significant; $p(\text{regions}) = 5.09 \times 10^{-8}$ and $p(\text{domains}) = 2.16 \times 10^{-6}$. Consistent with codon 5'-base invariance within amino acid families [7] and code domains (Fig. 1b), tRNA anticodon 3'-base invariance was significant within tree regions; $p = 1.04 \times 10^{-6}$ [8].

The correlations apparent in tRNA diversification, codon allocation, and amino acid pathway growth during code evolution imply tRNA cofactor/adaptor participation in pre-LCA amino acid synthesis coordinated new amino acid synthesis with codon recruitment.

3. Reconstruction of tRNA-dependent amino acid synthesis pathways

Biosynthesis reaction sequences [15] together with pre-LCA tRNA sequence identity and core structure [8, 9] have provided the framework for reconstruction of the pre-LCA tRNA-dependent amino acid synthesis pathways. Three precursors in amino acid biosynthesis were found to have initially shared an upstream precursor, reducing the number of amino acid families to two, in the pre-LCA era. Evidence of tRNA cofactor exchange is also shown here to resolve some anomalies apparent in tRNA specificity.

3.1. Upstream precursor

Alanine (Ala), valine (Val), and leucine (Leu) biosynthesis extends 1, 4, and 7 steps, respectively, from pyruvate (Pyr) [1]. Alanine and Val have IA-type tRNA, whereas Leu has a type-II tRNA [8, 9]. tRNA-$IA^{Ala}_{3'CGU}$ and tRNA-IA^{Asn}_{UUU} (later acquired by lysine [1, 2]) share type-IA tRNA [8, 9]. They also have pre-LCA identity of 11 quarts [8]; $p = 4^{-11} = 2.38 \times 10^{-7}$. An identity I_{ij} in the conserved imprint of pre-LCA tRNA species i and j is formally an expression of the binomial probability $p(x_{ij})$ that x_{ij} many (non-universal) sites were jointly conserved, by chance, in the consensus sequence of each tRNA from sources in each species domain; $I_{ij} = - \log_4 p(x_{ij})$ [8].

These similarities furnish evidence that tRNA-IA^{Ala} diversified from tRNA-IA^{Asn}. Path-distances notably place Asn in the first generation of coded amino acid [1, 2]. Asp-family amino acids threonine (Thr), isoleucine (Ile), methionine (Met), and lysine (Lys), likewise, possess type-IA tRNA with elevated pre-LCA identity versus tRNA-IA^{Asn}. By analogy with Asp-family amino acids, misacylation of a variant tRNA-IA^{Asn} by Asp initiated pre-LCA Ala synthesis.

This shifts the origin of Ala synthesis upstream from Pyr to OA (Fig. 2a). Indicative of masking by an attached tRNA cofactor, each reactant in the reconstructed tRNA-dependent Ala synthesis pathway, OA → Asp-tRNA → OA-tRNA → Pyr-tRNA → Ala-tRNA, contains a free α–carboxyl. Oxaloacteate and Pyr amination reactions combine to give Ala a path-distance of 2-steps. Path-distances for Val and Leu are also extended by 1-step, giving them a synthetic-order of 5- and 8-steps, respectively. Supernumerary Ala pathway reactions that predate amino acid synthesis and translation include: citrate cycle (CC) steps and tRNA acylation. Reversing the amination of Asp-tRNA to couple OA to a tRNA cofactor

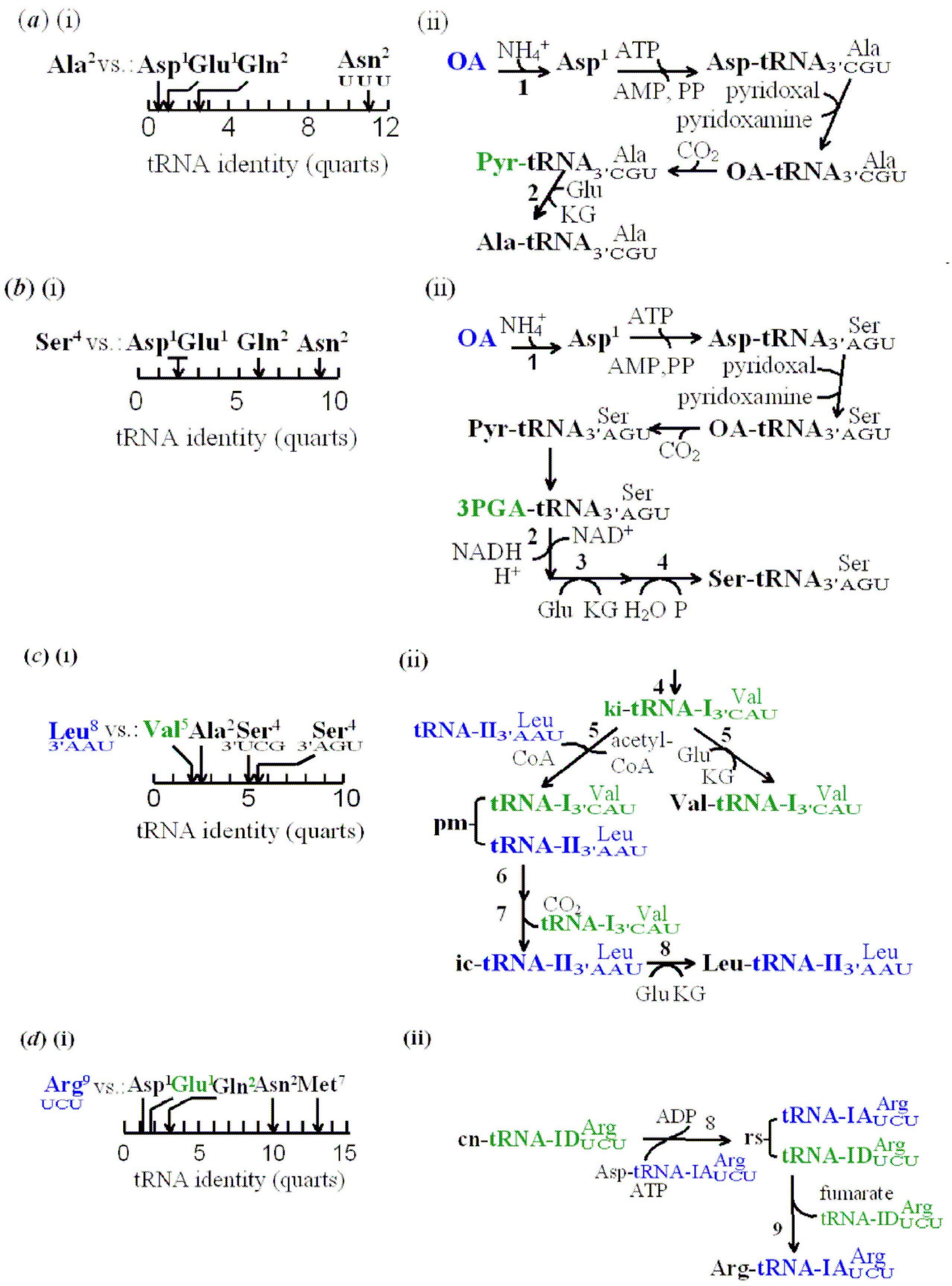

(a) (i)
Ala2 vs.: Asp1Glu1Gln2 Asn2
UUU
0 4 8 12
tRNA identity (quarts)
(ii)
OA NH4+ 1 Asp1 ATP AMP, PP Asp-tRNA3'CGU Ala
pyridoxal pyridoxamine
OA-tRNA3'CGU Ala
CO2
Pyr-tRNA3'CGU Ala
2 Glu KG
Ala-tRNA3'CGU Ala
(b) (i)
Ser4 vs.: Asp1Glu1 Gln2 Asn2
0 5 10
tRNA identity (quarts)
(ii)
OA NH4+ 1 Asp1 ATP AMP,PP Asp-tRNA3'AGU Ser
pyridoxal pyridoxamine
OA-tRNA3'AGU Ser
Pyr-tRNA3'AGU Ser
CO2
3PGA-tRNA3'AGU Ser
2 NAD+ NADH H+
3 4
Glu KG H2O P
Ser-tRNA3'AGU Ser
(c) (i)
Leu8 3'AAU vs.: Val5Ala2Ser4 3'UCG Ser4 3'AGU
0 5 10
tRNA identity (quarts)
(ii)
4
ki-tRNA-I3'CAU Val
tRNA-II3'AAU Leu
5 acetyl-CoA CoA
5 Glu KG
Val-tRNA-I3'CAU Val
pm tRNA-I3'CAU Val
tRNA-II3'AAU Leu
6
7 CO2 tRNA-I3'CAU Val
ic-tRNA-II3'AAU Leu
8 Glu KG
Leu-tRNA-II3'AAU Leu
(d) (i)
Arg9 UCU vs.: Asp1Glu1Gln2Asn2Met7
0 5 10 15
tRNA identity (quarts)
(ii)
cn-tRNA-IDUCU Arg
ADP 8
Asp-tRNA-IAUCU Arg
ATP
rs tRNA-IAUCU Arg
tRNA-IDUCU Arg
9 fumarate tRNA-IDUCU Arg
Arg-tRNA-IAUCU Arg

Figure 2. Reconstructed tRNA-dependent amino acid synthesis pathways illustrating upstream precursors and cofactor-exchange. (*a*) Identity (quarts) at conserved (non-universal) tRNA sites and shared tRNA core structure group indicate tRNA-IA$^{Ala}_{3'CGU}$ diversified from tRNA-IA$^{Asn}_{UUU}$ [8]. Initial misacylation of tRNA-IA$^{Ala}_{3'CGU}$ with Asp requires extension of the Ala pathway upstream from Pyr to OA. Central metabolism step OA → Pyr predates Pyr → Ala and code evolution, making it a supernumerary step to the time-order of Ala entry to code. (*b*) Pre-divergence tRNA identity (variable loop excluded) shows tRNA-IA^{Asn} to be ancestral to tRNA-II$^{Ser}_{3'AGU}$. Charging the Ser tRNA cofactor with precursor amino acid, Asp, requires an OA → Pyr → 3PG segment in the Ser pathway. (c) Leucine has type-II tRNA, whereas other Pyr derived amino acids, Ala^2 and Val^5, have type-IA tRNA. Dicarboxylated intermediate, α-isopropyl-malate (pm), at step-5 in Leu synthesis, initiated the switch in its tRNA cofactor. Pre-LCA identity reveals tRNA-II^{Leu} arose from tRNA-II^{Ser}. (d) Asp is last precursor in synthesis of Arg, which has Asn-like type-IA tRNA. Its initial precursor, Glu, misacylated a type-ID tRNA. A dicarboxylated intermediate, arginine-succinate (rs), at step 8 in Arg synthesis, enables the transition from type-ID to -IA tRNA cofactor.

Is also discounted. Non-enzymatic β-decarboxylation directly converts Asp to Ala is suggestive of a prebiotic path [16]. However, this route omits Pyr, decoupling pre-LCA Ala synthesis from its later path in biosynthesis, and for this reason it is discounted.

Pre-LCA identities indicate tRNA-IA^{Asn} was the source of tRNA specific for several other amino acids whose biosynthesis shows no direct association with Asn. They include Serine (Ser) and related amino acids Cys, Glycine (Gly), and Trptophan (Trp). tRNA-II$^{Ser}_{3'AGU}$ has an identity of 9.2 quarts (p = 2.9x10^{-6}) with tRNA-IA$^{Asn}_{UUG}$, variable loop excluded [8]. tRNA$^{Ser}_{3'UCG}$ has a lower identity (5.7 quarts, p = 3.7x10^{-4} [8]) with tRNA-IA$^{Asn}_{3'UUG}$, consistent with Ser acquiring codons AGY after UCN [17]. Acquisition of a type-II tRNA by Ser apparently occurred after Cys formation, since Cys retains an Asn-like type-IA' tRNA. These findings parallel those for pre-LCA Ala synthesis. Misacylation of tRNA$^{Ser}_{3'AGU}$ with Asp is, accordingly, inferred to have also initiated Ser synthesis (Fig. 2b). Central metabolism reactions upstream from 3PGA, involving phosphorylation, hydration, and isomerization, follow deamination of Asp-tRNA (Fig. 2b). These reactions predate synthesis of the first coded amino acid, so they do not contribute to Ser synthetic-order. Oxaloacetate amination adds 1-step to Ser biosynthesis, which extends 3 steps from 3-phospho-glycerate (3PGA) in the central trunk [15]. Thus, Ser has a synthetic-order of 4-steps [1, 2]. Cysteine and Gly form by 1-step extensions of the Ser^4 pathway, giving each a synthetic-order of 5-steps.

The Ser-family acquired Trp when indole combined with Ser^4 [15], putatively conveyed by a variant tRNA-IA'^{Ser} cofactor. Pre-LCA sequence identity [8] reveals tRNA$^{Trp}_{3'ACC}$ arose from tRNA$^{Ser}_{3'AGU}$; I = 7.0 quarts, p = 6.1x10^{-5}. tRNA-IA'$^{Trp}_{3'ACC}$ shares a type-IA' core group with tRNA-IA'$^{Cys}_{3'ACG.}$ They form a small code domain with contiguous codons UGG and UG^U_C (Fig. 1b).

tRNA cofactor/adaptors for 14 amino acids can be traced to tRNA-IA^{Asn} [8]. tRNA species for the remaining 5 amino acids convey precursor amino acids, Asp and Glu, plus Glu-family amino acids Gln, proline (Pro), and histidine (His). The large excess in Asp produced amino acids is interpreted in Table S1 (#42). Aspartate has a type-ID tRNA, with no significant identity for tRNA-IA^{Asn}; I = 1.5 quarts, p = 0.125 [8]. Pre-LCA tRNA-ID$^{Asp}_{3'CUG}$ and tRNA-ID$^{Glu}_{3'CUU}$ are however related; I = 7.0 quarts, p = 6.10x10^{-5}. Both also share a type-ID core structure [8]. tRNA sub-types of precursor and product amino acids in the Asp-family differ, in a departure from the precursor-product hypothesis [18, 19], suggesting product amino acids captured precursor tRNA.

Figure 2a,b show reconstructed tRNA-dependent pathways with a central metabolism segment. Pre-LCA tRNA in these pathways had elevated identity versus tRNA-IA^{Asn}, but convey amino acids biosynthetically unrelated to the Asp-family. These segments were seemingly lost, together with tRNA cofactors, during the protein take-over of amino acid synthesis. Their inclusion in pre-LCA amino acid synthesis pathways supports RNA initially coordinating and catalyzing CC and central trunk reactions (§7). The variety of amino acids from a single source (OA) was efficiently increased by branching these pathways at different points: Pyr to form alkyl chain amino acids, PEP aromatic amino acids, and 3PGA hydroxyl or sulfhydryl bearing amino acids. Path-selection by pre-LCA tRNA cofactor/adaptor molecules can be linked to path-identity elements (§6).

3.2. tRNA cofactor exchange

Valine and Leu pathways share the first three reactions downstream from Pyr [15]. They might be anticipated therefore to have related tRNA. Instead, they have type–IA' and type-II tRNA, respectively. With pre-LCA sequence identity of 1.8 quarts, p = 0.082 NS [8], tRNA-$IA'^{Val}_{3'CAU}$ and tRNA-$II^{Leu}_{3'AAU, 3'GAU}$ appear unrelated. Their codons (GU●, $^{C}_{U}$U●, 3'-bases suppressed) also have different 5'-bases, contrary to the 5'-base invariance rule [7]. A type-IA' → type-II tRNA exchange evidently occurred. Evidence for the exchange is found on the Leu path. Synthesis of di-carboxylated α-isopropyl-malate (pm), in the first step on the Leu pathway (step-5 from OA), could initiate the exchange (Fig. 2c). Decarboxylation of β-isopropyl-malate, 2-steps downstream from pm [15], produces mono-carboxylated α-keto-isocaproate (ic). This step would jettison the type-IA tRNA and complete the exchange. Subsequent amination of ic-tRNA-II^{Leu} produces Leu-tRNA-II^{Leu}.

tRNA^{Ser} is the only known source of type-II tRNA in the early stage-2 code. Pre-LCA tRNA-$II^{Leu}_{3'AAU}$ accordingly shows elevated identity with pre-LCA tRNA-$II^{Ser}_{3'AGU}$; I = 5.7 quarts, p = 3.70×10^{-4} [8]. Both tRNA also read nearest-neighbor codons, UC● and UU●, sharing a 5'-U. In addition to UCN and AGN, Ser apparently acquired 4-set UUN, and possibly CUN, during expansion of the stage-2 code. A type-IA → type-II tRNA exchange (§5.1) would facilitate assignment of multiple codon sets to Ser, as anticodon arm identity elements [20] could shift to the large variable loop of a type-II tRNA. Allotting multiple codon sets to Ser then becomes feasible. Appearance of tRNA-II^{Ser} early in stage-2, therefore, would have reduced the threat posed by an excess of unassigned/ nonsense triplets, which can block translation with lethal effect [21]. Reassigning codons UUN/CUN to Leu, a hydrophobic residue [22], coincided with emergence of membrane proteins, including the proteolipid subunit of [H^+]-ATPase late in stage-2 [23]. Non-polar residues Leu and Val, in addition, obtained nearest-neighbor codon 4-sets, $^{U}{}_{C}U$● and GU●; ΔF_T, Ser –0.6, Val –2.6, Leu –2.8 kcal/mol, where ΔF_T is the mean free energy change on transfer of a residue from an aqueous solution to a solvent with a dielectric constant of 2.0 [22].

Anomalies also arise in the synthesis and coding of Arg. Although Asp contributed a solitary N atom, Arg acquired an Asp-family type-IA tRNA [8] and 5'-A bearing codons (AGR). Other Arg codons have a 5'-C (CGN) linking the amino acid to its primary precursor, Glu^1. These incongruencies suggest a type-ID → type-IA tRNA exchange accompanied a Glu- → Asp-family transition during growth of the Arg pathway. Ornithine (Orn^6), and possibly citrulline (Cit^7), were likely incorporated into proteins prior to Arg^9 synthesis [24]. As a Glu-family member, Orn^6 had a type-ID tRNA cofactor/adaptor and codons in the CNN set. Path extension beyond Orn^6 accompanied transfer of the CGN 4-set to 'the end-product α-amino acid' [24]. Evidence of cofactor exchange is found beyond Orn^6 synthesis. A dicarboxylated intermediate, arginine-succinate (rs), is produced in the penultimate reaction in Arg synthesis. Figure 2d depicts Asp, conveyed by a $\text{tRNA-IA}^{Arg}{}_{UCU}$ cofactor, donating an N atom to Arg (guanidinium group). Arg-tRNA-IA^{Arg} putatively formed on lysing rs, with release of a fumarate-tRNA-ID complex. End-product α-amino acid coding [24] fits with subsequent formation of an isoacceptor, $\text{tRNA-IA}^{Arg}{}_{3'GCU}$, cognate with 4-set CGN.

Dicarboxylated-intermediates occur at steps 4 to 9 of Lys^{10} synthesis [15], following a reaction combining Pyr and aspartate-β-semialdehyde. Both Asp- and Pyr-family amino acids have IA-type tRNA. Pre-LCA sequence identities [8] show $\text{tRNA-IA}^{Lys}{}_{UUU}$ more closely resembles $\text{tRNA-IA}^{Asn}{}_{3'UUG}$ (I = 16.0 quarts, $p = 2.33\text{x}10^{-10}$) than $\text{tRNA-IA}^{Ala}{}_{3'CGU}$ (11.0 quarts, $p = 2.38\text{x}10^{-7}$). This makes tRNA cofactor exchange unlikely during Lys^{10} synthesis.

Retention of a second tRNA cofactor in Lys synthesis, suggested by the retention of dicarboxylated intermediates, possibly served to block incorporation of α–amino acid intermediates, diamino-L-pimelate and meso-diamino-L-pimelate.

4. Amino acid synthetic-order

Figure 3 shows the synthetic-order of code amino acids based on the number of reaction steps in reconstructed tRNA-dependent pathways. Seventeen amino acids retain their biosynthesis path-distances [1, 2, 8, 22, 25]. Alanine^{2}, Val^{5}, and Leu^{8} show a 1-step increase. Agreement between pre-LCA tRNA-dependent and biosynthesis path-distances reflects the highly conserved nature of these reaction sequences [26]. This invariance underlies the success of biosynthesis path-distances in unifying over twenty different code features [8]. Central metabolism segments connecting precursor OA to downstream branch-reactions at Pyr, PEP, and 3PGA contributed generic differences between amino acids from each branch-reaction (§5.1). The antiquity of central metabolism segments renders them supernumerary with respect to amino acid synthetic-order. Oxaloacetate amination added 1-step to biosynthesis pathways branching at Pyr, 3PGA, and PEP. To allow for the long association of the CC (source of over half amino acids) with protein synthesis, path-distances in biosynthesis pathways originating in the central trunk (3PGA, PEP) extended by 1-step [1].

Sequence identities between pre-LCA tRNA species indicate early amino acid synthesis pathways originated at either OA or KG. Aspartate^{1}, produced by amination of OA, was precursor to 14 amino acids and Glu^{1}, from KG, produced three (Fig. 3). Ribose-5-phosphate initiated His^{13} synthesis. Acquisition of a type-ID tRNA cognate with codon doublet, CAY, in a 4-set shared with Gln^{2}, added His^{13} to the Glu^{1} family. Glutamine^{2} donates an N atom to the His imidazole side-ring [15] and its tRNA cofactor is the likely source of tRNA-ID$^{His}_{GUG}$.

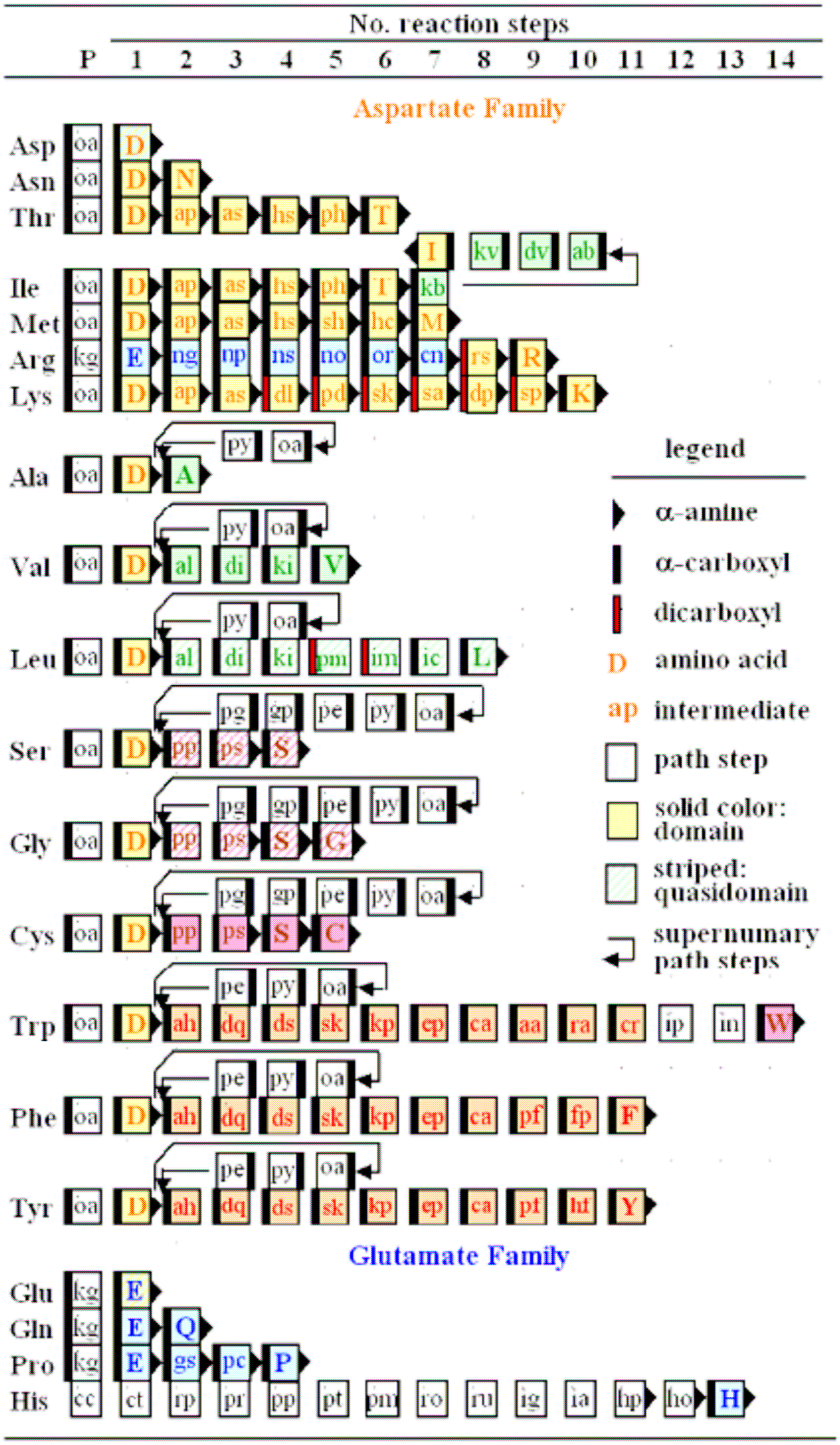

No. reaction steps
P 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Aspartate Family
Asp oa D
Asn oa D N
Thr oa D ap as hs ph T
I kv dv ab
Ile oa D ap as hs ph T kb
Met oa D ap as hs sh hc M
Arg kg E ng np ns no or cn rs R
Lys oa D ap as dl pd sk sa dp sp K
py oa
Ala oa D A
py oa
Val oa D al di ki V
py oa
Leu oa D al di ki pm im ic L
pg gp pe py oa
Ser oa D pp ps S
pg gp pe py oa
Gly oa D pp ps S G
pg gp pe py oa
Cys oa D pp ps S C
pe py oa
Trp oa D ah dq ds sk kp ep ca aa ra cr ip in W
pe py oa
Phe oa D ah dq ds sk kp ep ca pf fp F
pe py oa
Tyr oa D ah dq ds sk kp ep ca pt ht Y
Glutamate Family
Glu kg E
Gln kg E Q
Pro kg E gs pc P
His cc ct rp pr pp pt pm ro ru ig ia hp ho H
legend
α-amine
α-carboxyl
dicarboxyl
D amino acid
ap intermediate
path step
solid color: domain
striped: quasidomain
supernumary path steps

Figure 3. Amino acid path-distances in reconstructed tRNA-dependent synthesis pathways. An overbar indicates the number of reaction steps (synthetic-order) in each. Oxaloacetate (oa) is precursor (P) to fifteen amino acids, yielding a super-aspartate family. Ketoglutarate (kg) produced four amino acids in the glutamate family and one arose from ribose-5-phosphate (rp). Letter and background colors indicate a code domain or quasi-domain, as in Fig. 1b. Black letters on white background indicate a citrate cycle (cc) or central trunk (ct) metabolite, or an intermediate lacking a tRNA cofactor. Left-side bar signifies an α-carboxyl group; right-side triangle represents an α-amine. Loops contain supernumerary reactions in central metabolism segments. Three-letter amino acid abbreviations appear in the left-hand column. Upper-case, single-letter amino acid abbreviations appear within pathways. Lower-case, double-letter abbreviations identify non-amino-acid intermediates [15]**:** py, pyruvate, pe, phosphoenolpyruvate, gp, 2-phosphoglycerate, pg, 3-phosphoglycerate. **Thr** - ap, aspartyl-phosphate; as, aspartate-β-semialdehyde; hs, homoserine; ph, o-phospho-homo-serine. **Ile** - kb, α-keto-butyrate; ab, α-aceto-α-hydroxy-butyrate; dv, α,β-dihydroxy-iso-valerate; kv, α-keto-isovalerate. **Met** – sh, o-succinyl-homoserine; hc, homocysteine. **Arg** - ng, N-acetyl-glutamate; np, N-acetyl-glutamate-phosphate; ns, N-acetyl-glutamate-γ-semi-aldehyde; no, N-acetyl-ornithine; or, ornithine; cn, citrulline; rs, arginine-succinate. **Lys** - dl, α,β-dihydro-picolineate; pd, Δ^1-piperdiene-2,6-dicarboxylate; sk, N-succinyl-ε-keto- α-amino-pimelate; sa, N-succinyl- α,ε-diamino-pimelate; dp, α,ε-diamino-L-pimelate; sp, meso- αε-diamino-pimelate. **Ala** - nd, Glu amine-donor. **Val** - al, α-aceto-lactate; dl, α,β-dihydroxy-iso-valerate; kl, α-keto-isovalerate. **Leu** - pm, α-isopropyl-malate; im, β-isopropyl-malate; ic, α-keto-isocaproate. **Ser** – op, phospho-hydroxypyruvate; ps, phospho-serine. **Trp** - ah, β-deoxy-arabino-heptulosonate-7-phosohate; dq, 5-dehydroquinate; ds, 5-dehydro-shikimate; sk, shikimate; kp, shikimate-5-phisohate; ps, 3-enolpyruvyl-shikimate-5-phosphate; ca, chorismate; aa, anthranilate; ra, N-phospho-ribosyl-anthranilate; cr, 1-(o-carboxyphenyl-amino)-1'-deoxyribulose-5-phosphate; ip, indole-3-glycerol-phosphate; in, indole. **Phe** - pf, prephenate; fp, phenyl-pyruvate. **Tyr** - hf, p-hydroxy-phenyl-pyruvate. **Pro** - gs, glutamate-γ-semialdehyde; pc, Δ1-pyrroline-5'-carboxylate. **His** - pp, phosphatidyl-ribosyl-pyro-phosphate; pt, phospho-ribosyl-adenosine-triphosphate; rm, phospho-ribosyl-adenosine-monophosphate; ro, phospho-ribosyl-formimino-amino-imidazole-carboxamide-ribose-phosphate; ru, phospho-ribulosyl-formimino-amino-imidazole-carboxamide-ribose phosphate; ig, erythro-imidazole-gylcerol-phosphate; ia, imidazole-acetol-phosphate; hp, histidinol-phosphate; ho, histidinol.

tRNA for Asp-family amino acids (Fig.1c) form a tree with tRNA-IA^{Asn} at its root [8]. Variant tRNA-IA^{Asn} served as cofactors in the synthesis of new amino acids. Misacylation of tRNA-IA^{Asn} by Asp^{1}, in the first code [1, 2, 8, 22, 25], thus provided a blueprint for code expansion. Restriction of direct (tRNA-cofactor-free) amino acid synthesis to precursors Asp^{1} and Glu^{1}, and the failure of tRNA-ID^{Asp} and tRNA-ID^{Glu} to diversify [24], conform with extensive pre-LCA reliance on tRNA-dependent amino acid synthesis. tRNA cofactor/ adaptor participation in matching amino acids with their codons (§5.1) undoubtedly conferred an initial advantage on indirect amino acid synthesis.

Clusters of Asp- and Glu-family amino acids with ANN and CNN set triplets read by type-IA and type-ID tRNA, respectively, provide further evidence of early tRNA participation in amino acid synthesis and its impact on formation of the standard code (Fig. 1b). Fragmentation of the large Asp family accompanied by elimination of both tRNA cofactors and upstream reactions from Pyr, PEP, and 3PGA, is linked to the protein takeover of these pathways [8]. Three new amino acid families arose in the transition to direct amino acid synthesis: (Ala^{2}, Val^{5}, Leu^{8}), (Phe^{11}, Tyr^{11}), and (Ser^{4}, Cys^{5}, Gly^{5}, Trp^{14}).

5. Imprint of amino acid synthetic-order on genetic code and early proteins

Code and early protein evolution is interpreted here using amino acid synthetic-order in reconstructed tRNA-dependent pathways (§4). The code is revealed to have formed in the three stages. Generically different amino acids entered the code in each stage, indicative of fundamental changes in the direction of protein evolution during code formation. In stage-1 and -2, coding capacity achieved or slightly surpassed its upper limit by retaining some initial ambiguity, while the stage-3 code (standard code) froze before its theoretical limit.

5.1. Genetic code

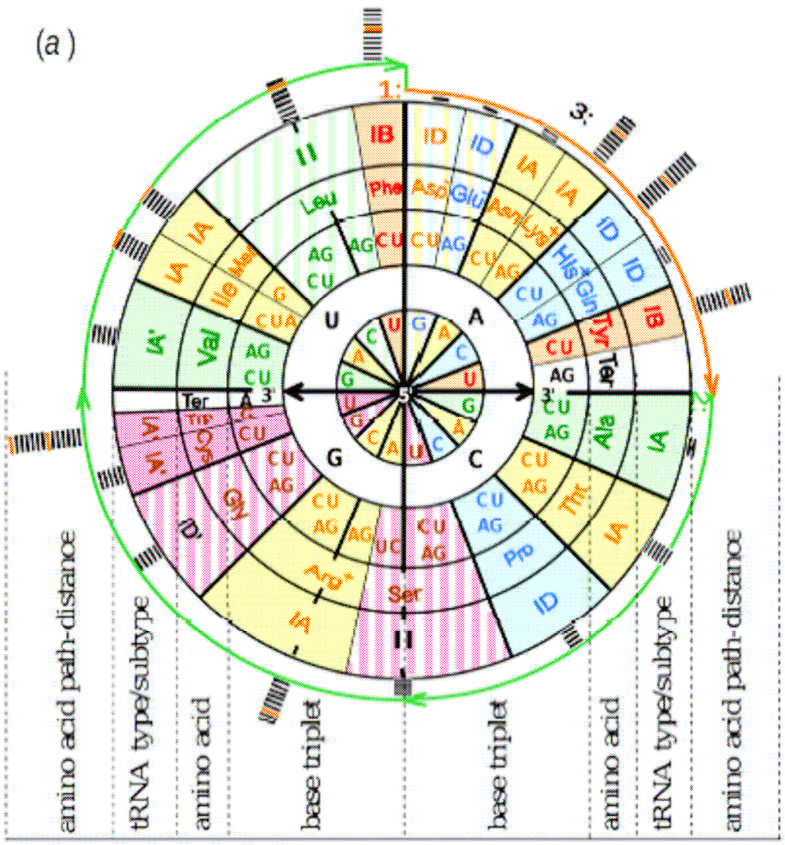
Path-distance evidence established the genetic code evolved by successively recruiting the codon 5'-, mid-, and 3'-site (Fig. 4). In the first code [1, 2], 16 A-quadrant (XAN) triplets coded for 2 diacid/amide amino acid pairs, Asp^1/Asn^2 and Glu^1/Gln^2, formed on 1-2 step paths, plus a chain termination (Ter)/STOP signal. Back-tracking from the A-rich triplets of the first code points to pre-code translation on a poly(A) strand producing random sequence poly(Asp,Glu,Asn,Gln) [1]. With a fixed mid-A and degenerate 3'-site, coding specificity resided in the codon 5'-base. The four XAN 4-sets could code for no more than four amino acids. All 16 triplets were 'sense' codons. Consequently, the risk of a 'nonsense' (unassigned) triplet, which could block translation [21], was limited to mid-base substitutions – any mutation at one-of-three codon sites. In a less compact code, this risk rises. A random codon distribution, for example, has a three-in-four risk ((64 -16)/64, nonsense/total triplets) [1]. Codon mid-base ambiguity during translation [27] of NAN triplets by tRNA species with a 'wobble' site U34 (Fig. 4) would allow translation directed by the small NH_4^+ Fixers Code to read-through an unassigned/nonsense triplet.

With four amino acids plus a STOP signal, the first code exceeded the coding limit of a single codon site. AAN, CAN, and UAN coded for Asn^2, Gln^2, and Ter, respectively; long-path (stage-3) amino acids, Lys^{10}, Tyr^{11}, His^{13}, being disregarded in the early code (Fig. 4a). It appears GAN triplets jointly and ambiguously coded for Asp^1 and Glu^1. First generation amino acids, Asp^1, Glu^1, Asn^2, Gln^2, facilitated the flow of N atoms from NH_4^+ to amino acids, nucleic acids, and coenzymes [1]. This places the origin of proteins within a primal NH_4^+ fixing/distribution system.

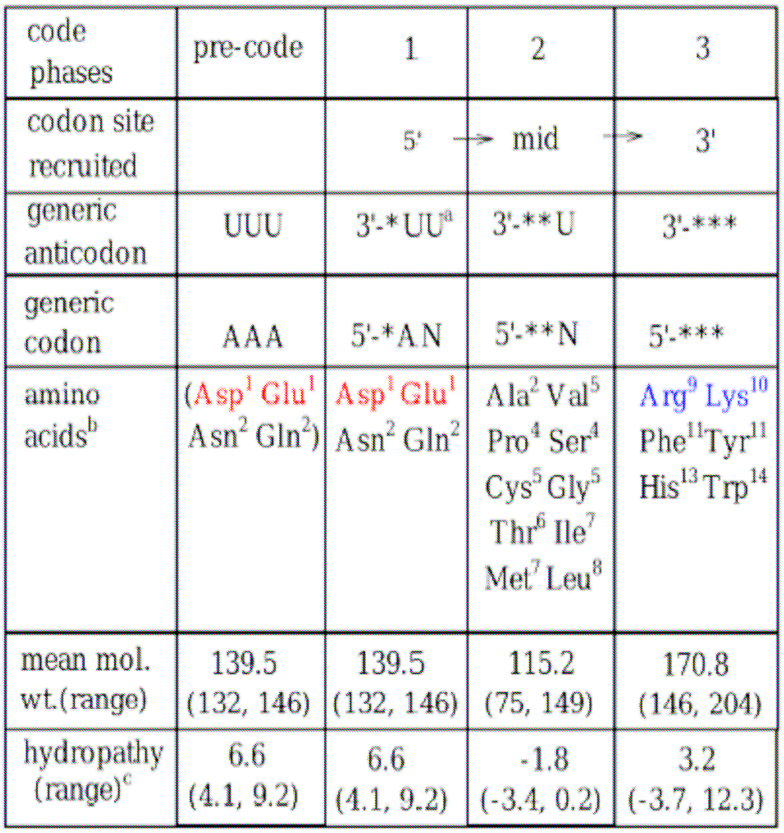
Class-I and -II aminoacyl-tRNA-synthetase enzymes act on Glu^1/Gln^2 and Asp^1/Asn^2, respectively [8]. This links the duality of the first generation of coded amino acids to the two-fold division among synthetases. Class I and II synthetase enzymes it appears from this had their origin in ribozymal antecedents that specifically acylated tRNA adaptors with Glu^1 or Asp^1 and misacylated the corresponding amide amino acid cofactor/adaptor in tRNA-dependent Gln^2 and Asn^2 synthesis [8].

Expansion beyond the NH_4^+ Fixers Code led to assignment of twelve codon 4-sets in the C-, G-, U-quadrants to ten mainly hydrophobic alkyl-chain amino acids and a STOP signal: Ala^2, Pro^4, Ser^4, Cys^5, Gly^5, Val^5, Thr^6, Ile^7, Met^7, Leu^8. Figure 4a shows code expansion during codon mid-base recruitment proceeded in the direction: (NAN) → NCN → NGN → NUN. Amino acid mean path-distance increased between quadrants, (1.5) → 4.0 → 4.7 → 7.0 steps (§5), consistent with quadrant-by-quadrant growth. This growth pattern preserved code compactness during stage-2 expansion. Increased residue hydrophobicity (ΔF_T) accompanied code expansion: (6.6) → -0.8 → -1.2 → -2.9 kcal/mol per quadrant. Synthesis of increasingly hydrophobic proteins thus accompanied code expansion [1]. Pre-LCA protein phylogenetics placed appearance of the first membrane proteins at late in code expansion [22]. Recruitment of AUN by an α-amino acid intermediate of Met^7 could have provided a START signal required for assembly of an ordered residue sequence on expansion from the first code.

(*a*)



(*b*)



code phases	pre-code	1	2	3
codon site recruited		5'	→ mid	→ 3'
generic anticodon	UUU	3'-*UU[a]	3'-**U	3'-***
generic codon	AAA	5'-*AN	5'-**N	5'-***
amino acids[b]	(Asp^{1} Glu^{1} Asn^{2} Gln^{2})	Asp^{1} Glu^{1} Asn^{2} Gln^{2}	Ala^{2} Val^{5} Pro^{4} Ser^{4} Cys^{5} Gly^{5} Thr^{6} Ile^{7} Met^{7} Leu^{8}	Arg^{9} Lys^{10} Phe^{11}Tyr^{11} His^{13}Trp^{14}
mean mol. wt.(range)	139.5 (132, 146)	139.5 (132, 146)	115.2 (75, 149)	170.8 (146, 204)
hydropathy (range)[c]	6.6 (4.1, 9.2)	6.6 (4.1, 9.2)	-1.8 (-3.4, 0.2)	3.2 (-3.7, 12.3)

Figure 4. Main stages of genetic code formation revealed by amino acid synthetic-order in reconstructed tRNA-dependent pathways. (a) Orange arrow marks stage-1. Four NH_4^+ fixer/donor amino acids, synthesized on 1-2 step paths, with A-quadrant codons. Green arrow designates stage-2. Ten mainly alkyl-chain amino acids, with path-distances (2-8 steps) increasing progressively in quadrants C, G, and U, from 4.0 ± 0.82 (mean± s.e.m.) to 4.7 ± 0.33 and 6.8 ± 0.63 steps, respectively, during code expansion. In stage-3, six post-expansion amino acids, 11.3 ± 0.76 steps, with large basic and aromatic side-chains captured 5 codon doublets and a single codon in error-prone triplet 4-sets, and 4-set CGN. Bar stacks depict path-distance of each amino acid, 1 bar/reaction step; an orange bar marks a 7-step span. Background colors designate code domains, as in Fig. 1b. (b) Codon sites were recruited in a 5'→3' direction (3'→5' anticodon sites), with triplet AAA (anticodon, UUU) used in pre-code.translation. Amino acids at each stage differ conspicuously in charge, molecular weight, and hydropathy. [a] X denotes coding site; [b] parenthesis, random residues; red, acidic residues; blue, basic; superscript, path-length. [c] residue transfer free energy (kcal/mol) [23].

Combining stage-2 additions to twelve 4-sets, including Orn^6 [24], together with the NH_4^+ fixers means the Stage-2 Code encoded 15 amino acids and a STOP signal. This equals the 16 'letter' limit attainable with two coding sites: 4 x 4. Further expansion of the code necessitated recruiting the codon 3'-site. An error suppression mechanism facilitated this. Subdivision of error-prone 4-sets [27] minimized codon mid-base misreading and allowed reassignment of a doublet to a stage-3 amino acid. Limiting the tRNA reading range to a natural doublet (Table S1, feature 3), by modifying or substituting U34, freed one doublet for reassignment. Post-expansion addition of six large basic and aromatic amino acids, Arg^9, Lys^{10}, Phe^{11}, Tyr^{11}, His^{13}, Trp^{14}, followed, increasing the amino acid alphabet to 20 amino acids, plus a STOP signal. All six long-path (> 8 steps) amino acids acquired codons in a subdivided 4-set shared with a short-path (≤ 8 steps) amino acid or STOP signal [8].

Arginine^9 also has 4-set CGN, attributed to transfer from Orn^6 [24], as required by the 'end-product α-amino acid' transfer rule following path extension (§3.2).

Six 4-sets in the standard code are split into natural doublets ($XX^{Y}{}_{R}$; R, purine, Y, pyrimidine) [28] and two, UGN, AUN, contain single codons assigned to Trp^{14} (UGG),.Ter (UGA), and Met^7 (AUG); with $AU^{Y}{}_{A}$ coding for Ile^7. Subdivision of all 16 XXR doublets to singles ($XX^{A}{}_{G}$) while retaining 3'-pyrimidine doublets, XXY [29], could expand the code to 47 amino acids and a STOP signal. As genome size increased, the increasing risk of a lethal substitution constrained code growth [30, 31]. Code structure suggests a second constraint on code growth. Splitting half the sixteen 4-sets in the stage –2 code into doublets increases coding capacity to 23 amino acids plus a STOP signal (eight 4-sets and sixteen natural doublets), close to the size of the standard code. Subdivision of eight codon 4-sets, as noted, is attributed to suppression of a translation reading error [27]. Expansion of the amino acid alphabet to 21 amino acids (with N-formyl-Met^8) and a STOP signal resulted. The resistance of GCN, GGN, GUN, ACN, CCN, UCN, CGN, CUN to subdivision, conversely, limited the coding capacity of each to one amino acid per 4-set. Translation fidelity and genome size thus combined to freeze the stage-3 code at 21 amino acids and a STOP signal.

Codon mid-base in the stage-2 code correlates with amino acids path-distance (≤ 8-step paths) [1, 2]. Code expansion from the NH_4^+ Fixers Code (Fig. 4a) consequently occurred by stepwise recruitment of triplets, through successive mid-base substitutions. This pattern of code growth preserved code compactness, reducing the risk of mutation to an unassigned/nonsense triplet. All sixteen codon 4-sets were allocated by completion of code expansion, given CGN coded for Orn^6 (§3.2). Further amino acid additions, therefore, required recruiting an available doublet or single in subdivided, error-prone 4-sets [25]. The ordered expansion in stage-2 ceased, as codon availability determined the distribution of long-path (9 – 14 steps) amino acids in the stage-3 code. Subdivision of 4-sets GAN and AUN assigned to NH_4^+ fixers, Asp^1/Glu^1, and expansion-stage amino acids, Ile^7/Met^7, also likely occurred in stage-3.

Successive mid-base substitutions in the anticodon of tRNA species originating from a common ancestor accounts for invariance of the codon 5'-base among pre-stage-3 same-family amino acids [8]. Triplets coding for stage-2 Asp-family amino acids Thr^6, Ile^7, and Met^7 illustrate this, as they share a 5'-A with stage-1 Asn^2 codons (Fig. 1b). Elevated pre-LCA identities furnish evidence their tRNA diversified from tRNA-IA^{Asn} [8]. Stage-3 amino acids also display codon 5'-base invariance with synthetically related amino acids. Thus, Asp-family members, Arg^9 and Lys^{10}, share codons from the ANN set with all four pre-stage-3 members, Asn^2, Thr^6, Met^7, Ile^7. Furthermore, they have type-IA tRNA related to tRNA-IA^{Asn} [8]. Codon 5'-base invariance among same-family amino acids [7] thus provides evidence for tRNA participation in amino acid synthesis throughout code formation. Occurrence of 5'-base invariance among stage-3 amino acids, after triplet recruitment by successive mid-base substitutions had ceased, points to a source of wider scope. This points to synthesome specificity for tRNA cofactor/adaptors participating in same-family amino acid synthesis (§6).

5.2 Early proteins

Residue sequences in proteins with known function and structure provide evidence of an amino acid potential for promoting, or suppressing, catalysis and secondary structure [32-

34]. An estimate of when proteins first acquired each feature is given in this section in relation to the stages identified in code formation [1, 2, 8].

Table 1 shows amino acids with significant potential for catalysis and α-helix, β-sheet, and β-turn formation at each stage in code evolution (Fig. 4). Amino acid potentials are expressed as the logarithm of the probability of a 2 x 2 contingency table portraying the association of a given residue with a specified protein feature versus all other residues. A protein assembled from a stage-1 residue profile, comprising the NH_4^+ fixers, could clearly form an α-helix with Glu^1 residues. Flanking the α-helix with anti-helical amide-residues could localize the helical segment within a protein. Stage–1 proteins were also capable of catalytic activity. Four stage-1 residues in the set, (Asn, Asp)-Asp-(Asp, Asn)-Asn, could form a β-turn. This motif requires only a few simple amino acids, prompting Jurka and Smith [35] to predict it arose early in protein evolution.

β-Sheets appeared first in proteins with stage-2 residues (Table 1). Anti-parallel β-sheets containing Ala^2 and Val^5, linked by β-turns with Pro^4, Ser^4, and Gly^5, and stage-1 residues Asp^1 and Asn^2 could form in proteins early in stage-2. Addition of increasingly hydrophobic amino acids during stage-2 (§5.1, 7) provides evidence of protein evolution toward a hydrophobic attractor [1, 2]. Proteins that partition with the cell membrane had emerged by advanced stages of code expansion [22]. Code evolution in stage-2 appears directed principally toward structural protein formation. Catalytic and structural amino acid potentials follow different timelines during code evolution. Apart from Cys^5, no stage-2 amino acid, for example, exhibits catalytic potential comparable to Asp^1 and Glu^1 in stage-1, or His^{13} and Arg^9 in stage-3. Seven of ten stage-2 amino acids actually have anti-catalytic potentials

Table 1. Amino acid potentials for catalysis and structural features at each stage of code formation. Values are -log *p* for pro-potentials and log *p* for anti-potentials, where *p* is the χ^2 probability ($0 < p < 1$) for residue frequency at sites with a designated attribute versus all other amino acids in a 2 x 2 contingency table. Values with $p < 1 \times 10^{-2}$ are listed. Parentheses indicate a sub-significant pro-amino acid. Catalytic potential sampled 191 proteins [32], α-helix and β-sheet structures 279 [33], and β-turns 59 proteins [34]. Frequencies and probabilities related to potentials appear in Table S2.

protein feature	**stage of code formation**					
	1		**2**		**3**	
	pro	**anti**	**pro**	**anti**	**pro**	**anti**
catalysis	Asp$^-$ 43.8 Glu$^-$ 12.4	Gln -2.0	Cys 16.0	Leu -19.2 Ala -18.1 Val -14.5 Ile -13.1 Pro -9.0 Gly -7.1 Met - 5.6	His 189.0 Arg$^+$ 26.0 Lys 5.1 Tyr 4.9	Phe -3.9
α-helix	Glu$^-$ 43.2 Gln 6.9	Asn -16.5	Ala 113.4 Leu 26.3 Met 7.1	Gly -103.4 Pro - 18.7 Thr - 16.0 Cys -7.3 Ser -5.3 Val -2.8	Arg$^+$ 11.2 Lys$^+$ 8.6	Tyr -6.0 His -4.8 Phe -3.5
β-turn : *i*	Asn 13.7 Asp 7.5			Val -3.1		
i+1	Asp$^-$ 2.3		Pro 6.1 Ser 5.6	Gly -2.0 Val -2.5		
i+2	Asp$^-$ 10.2 Asn 2.5		Ser 2.5	Pro -2.8 Val -2.4		
i+3	(Asn 0.6)		Gly 14.7	Pro -3.3	Trp 2.9	
β-sheet		Asp$^-$ -34.5 Glu$^-$ -19.0 Asn -8.6 Gln -2.7	Val 219.7 Ile 105.8 Thr 63.3 Cys 15.7	Pro -207.7 Gly -171.0 Ala -31.8	Tyr 41.1 Phe 35.4 Trp 4.3 His 2.1	Lys -2.3
		α-helix β-turn pro-enzyme		β-sheet		acid-base catalysis

(Table 1). Consequently, enzyme synthesis can be discounted as a significant determinant in shaping the amino acid alphabet of the stage-2 code. The possibility that expression of catalytic potential might require a nearly complete code can be discounted [36]. Ribozymes plainly remained the principal catalysts [35] during stage-2 of code formation.

Four stage-3 amino acids display significant catalytic potential (Table 1). Addition of the first basic amino acids, Arg^{9}, Lys^{10}, to the stage-3 code opened the way for acid-base catalysis. Late entry of basic amino acids into the code contrasts with occurrence of diacid amino acids in the first code. This provides evidence for the initial importance of charge attraction to a cationic mineral surface. Incorporation of His^{13} with its purine-like imidazole ring, in the final stage of code formation, undoubtedly aided the protein takeover of ribozyme catalyzed reactions.

6. tRNA cofactor mediated coding specificity

Widespread tRNA cofactor participation in early amino acid synthesis {§2) implies ribozymal synthetases [38] played no initial role in determining coding specificity, beyond charging tRNA cognate with Asp^{1} or Glu^{1} (§3.1) and mischarging tRNA cofactors with precursor. Coding specificity during code formation, consequently, relied on some alternative RNA mechanism [39]. Synthetase emergence accompanied tRNA cofactor replacement, raising the possibility tRNA contributed to amino acid selection and its role became redundant. Prokaryote synthesis of Asn^{2} and Gln^{2} [11] on tRNA-dependent pathways reveals that tRNA path-identity elements recruit catalysts to synthesize its cognate amino acid. tRNA-dependent amino acid synthesis, furthermore, takes place in a ribonucleoprotein particle equipped with multiple catalysts [40].

tRNA grafts [11] show conversion of Asp-$tRNA^{Asn}$ → Asn-$tRNA^{Asn}$ by Bacteria transamidase GatCAB requires recognition of a single acceptor-stem bp, U1:A72 (Fig. 5a). In Archaea, variable-loop bases, G46, U47, route Asp-$tRNA^{Asn}$ to Asn-$tRNA^{Asn}$ formation. Precursor adaptor, $tRNA^{Asp}$, contains anti-identity elements G1:C72, U20 in Bacteria and A46, Δ47 (Δ, deletion) in Archaea. When grafted into Asp-$tRNA^{Asn}$, they block amidation. In marked contrast to synthetase recognition [20], the tRNA anticodon is superfluous to GatCAB catalyzed transamidation. Since recognition of both amino acid and anticodon ensures coding specificity, anticodon irrelevance in transamidation suggests the Asn pathway contains a second recognition step. It arises on acylation of $tRNA^{Asn}$ with Asp, by a synthetase specific for both $tRNA^{Asp}$ and $tRNA^{Asn}$. Conversion of Asp-$tRNA^{Asn}$ to Asn-$tRNA^{Asn}$ proceeds via phosphorylation of Asp by a GatCAB kinase and subsequent amidation by a transamidmiase. $Asparagine^2$ synthesis occurs within a ribonucleoprotein particle (‘synthesome’), containing 2 synthetase molecules, 2 GatCAB, and 4 $tRNA^{Asn}$ [40]. Two $tRNA^{Asn}$ are charged and released, and 2 uncharged $tRNA^{Asn}$ provide a structural scaffold. Figure 5b depicts tRNA-dependent synthesis of Asn^2 and Gln^2 within a synthesome. Prokaryote GatCAB catalyzes amidation of both Asp-$tRNA^{Asn}$ and Glu-$tRNA^{Gln}$. Generic-substrate amidation is coupled with family-specific acylation, as the synthetases charge tRNA cognate with synthetically related amino acids, attaching Asp^1 to $tRNA^{Asp}$ and $tRNA^{Asn}$, and Glu^1 to $tRNA^{Glu}$ and $tRNA^{Gln}$.

Path-identity elements similar to those that route Asp-$tRNA^{Asn}$ to Asn^2 synthesis [11], provide a blueprint for a general mechanism of coding amino acids formed on pre-LCA tRNA-dependent pathways [10]. Synthetase enzyme recognition of an amino acid and its tRNA adaptor, accompanied by nearly complete elimination of tRNA cofactors, obscured

(*a*)

GatCAB	tRNA	Asp-tRNA^{Asp}		Asp-tRNA^{Asn}	
source	site	0	-	+	0
Bacteria	acceptor stem		G1:C72	U1:A72	
	D-loop		U20s	-	
Archaea	V-loop		A46, Δ47	G46, U47	
	anti-codon	3'CUG			3'UUG

(*b*)

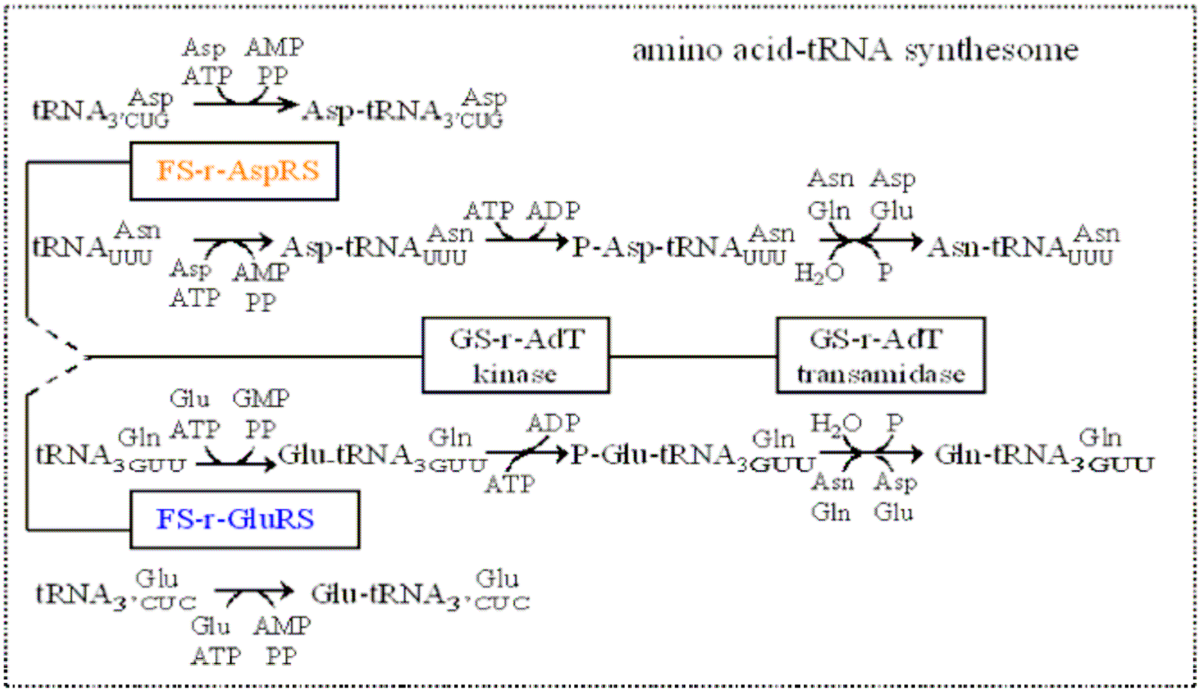


Figure 5. Catalyst recruitment by identity elements in tRNA cofactors during amino acid synthesis illustrated by Asn and Gln formation. (*a*) Site of tRNA identity elements in recruitment of prokaryote transamidase, GatCAB, during homeotopic amidation of aspartate attached to $tRNA^{Asn}$: +, pro-reaction and, -, anti-reaction identity elements; 0, neutral features; s, supernumerary base; and Δ, deletion**.** Based on results of tRNA graft experiments [11]. (*b*) tRNA cofactor directed amidation of precursor diacid amino acids within an amide amino acid-tRNA synthesome, depicted as containing either Asp-family or Glu-family syntherases, and kinases and amidotransferases with generic specificity for amide amino acid tRNA. Aspartate and Glu charge specific bifunctional tRNA species, which serve as cofactors in the synthesis of product amino acids, Asn and Gln, and as adaptors in their translation. FS-r-AspRS, ribozymal antecedent of aspartyl-tRNA synthetase specific for Asp-family amino acids. FS-r-GluRS, ribozymal synthetase specific for Glu family amino acids. GS-r-AdT ribozymal amidotransferase with generic specificity for amino acyl-$tRNA^{amide\ amino\ acid}$.

how coding specificity was achieved by the RNA code. Progress in identifying conserved features of code structure, pre-LCA tRNA sequences, and prokaryote tRNA-dependent pathways helped provide a solution (§5.1).

The scope of pre-LCA tRNA-dependent amino acid synthesis (§2) suggests transamidosome-like synthesomes [40] were widely distributed in the pre-LCA era. By combining Thr^6 and Val^5 pathways, Ile^7 synthesis illustrates this (Fig. 3). An independent enzyme catalyzes each step in Val biosynthesis: acetolactate synthase, acetohydroxy acid isomero-reductase, dihydroxy acid dehydratase, and valine aminotransferase [15]. Splicing the whole 4-step Val reaction sequence onto the Thr^6 pathway, however, conforms with the transfer of a cassette of ribozymes. This fits with synthesome participation in pre-LCA Val synthesis. The branching pattern of same-family amino acid pathways [1], where individual amino acid paths diverge from a common reaction sequence, also conforms with early reliance on a central synthesome.

7. Origin of the genetic code and its implications

Establishing the origin of the genetic code has proved to be a long and arduous undertaking. To a large extent this reflects code complexity (Table S1). Reconstructing the path of code evolution, more than 3.5 billion years distant [41- 43], is seen to require more than a superficially haphazard set of 64 codon assignments [44]. In particular, tRNA proved to be a key player in code formation, coordinating synthesis of new amino acids with triplet recruitment (§2, 3).

Almost all attempts to explain the origin of the genetic code have focused on identifying the source of a single feature in its many-faceted structure [2]. Trifonov took a broader approach, evaluating the consensus order of amino acid entry into the code among forty different scenarios of code formation [45]. Abiogenically synthesized amino acids [46] emerged as most likely to have formed the first code: Gly, Ala, Asp, Glu, Pro, Ser, Leu, Thr. A consensus of many different scenarios might be anticipated to approximate, to some degree, the actual time-order of amino acid addition to the code. In the absence of the actual sequence, however, the margin of error remains indeterminate. The present reconstruction based on code structure, amino acid synthetic-order, and sequence identity in pre-LCA tRNA species, not surprisingly, places a different set of amino acids in the first code: Asp^1, Glu^1, Asn^2, Gln^2 (Fig. 4). In any realistic scenario, the path leading to the standard code had many branches. Since the standard code alone survives from the pre-LCA era, matching a proposed time-order with known features of code structure [2] is the only apparent means of validation. A time-order that closely approximates the actual path of code evolution could be anticipated, in addition, to reveal formerly unnoticed features of code structure. Half the fifty-two code features listed in Table S1 were notably uncovered by the path-distance principle.

The transition to ordered polypeptide synthesis, on formation of the NH_4^+ Fixers Code, allowed optimization of acidic/amide residue mole ratios [1]. Appearance of simple folded proto-enzymes could follow recruitment of AUN as a START signal by a Met^7 intermediate (§6). Aspartate^1, Glu^1, Asn^2, and Gln^2 are hydrophilic homologues with a mean residue ΔF_T of 6.58 ± 1.25 (m ± s.e.m.) and Sneath homology coefficient of 0.64 ± 0.05 per amino acid pair (n(aa) = 4, n(aa pairs) = 6), consistent with elevated substitution rates [47]. Values of only 1.38 ± 1.09 kcal/mol and 0.38 ± 0.03 per pair (n(aa) = 20, n(aa pairs) = 190), respectively, apply among all amino acids [1, 8, 22]. Minimizing substitution rates between residues of different polarity [4-6] had no apparent role in forming a code of homologues.

The 7-residue cluster of polar amino acids at NAN in the standard code (Fig. 1b) formed in two stages. NAN set latecomers Lys^{10}, Tyr^{11}, and His^{13} acquired doublets AAR, UAY, and CAY, following stage-3 subdivision of the respective 4-sets (Fig. 4a). Like the stage-1 NH_4^+ fixers, they have positive residue ΔF_T values: 8.8, 0.7, and 3.0 kcal/mol, respectively [23]. They lowered mean Sneath homology in the NAN set cluster, however, from 0.64 ± 0.05 (NH_4^+ fixers) to 0.36 ± 0.03 (n(pairs) = 15, over the three additions).

Triplet availability thus contributed to formation of the NAN cluster. Path-distances of amino acids encoded by 5'-base sets ANN, UNN, and CNN reveal doublets AAR, UAY, and CAY were the last triplets available for allocation to Lys^{10}, Tyr^{11}, and His^{13}, respectively (Figs. 1b, 4a). Although Trp^{14} acquired UGG after Tyr^{11} obtained $UA^U{}_C$ in the UNN set, $U^U{}_A{}^U{}_C$ triplets are not contiguous with $UG^A{}_G$, placing the latter outside the Phe/Tyr domain and excluding them from coding for Tyr^{11} [8]. The catalytic potential of Lys^{10}, Tyr^{11}, and His^{13}

(Table 1) undoubtedly contributed to their incorporation into proteins. Stage-3 subdivision of NAN 4-sets [27], the direction of protein evolution in stage-1 and –3 of code evolution (Fig. 4), and triplet availability thus contributed to formation of the polar-residue cluster.

Subdivided 4-sets $AA^{R}{}_{Y}$ and $CA^{Y}{}_{R}$, and nearest-neighbor triplets $U^{A}{}_{U}Y$, respectively, code for same-family amino acid pairs Lys^{10}/Asn^{2}, His^{13}/Gln^{2}, and Tyr^{11}/Phe^{11}. Cofactor/adaptors pairs tRNA-$IA^{Lys}{}_{UUU}$/tRNA-$IA^{Asn}{}_{3'UUG}$, tRNA-$ID^{His}{}_{GUG}$/tRNA-$ID^{Gln}{}_{3'GUU}$, and tRNA-$IB^{Tyr}{}_{3'AUG}$/tRNA-$IB^{Phe}{}_{3'AAG,}$ contain the same core group and, in two pairs, elevated pre-LCA sequence identity [8]; 16.0 ($p = 2.3\text{x}10^{-10}$), 1.4 ($p = 0.14$), and 10.0 ($p = 9.5\text{x}10^{-7}$) quarts, respectively. Tyrosine^{11} and Phe^{11} both form on 11-step paths, but pre-LCA identity [8] is higher between tRNA-A^{Phe} and ancestral tRNA-$IA^{Met}{}_{3'UAC}$: tRNA-$IB^{Phe}{}_{3'AAG}$, 16.0 and tRNA-$IB^{Tyr}{}_{3'AUG}$, 9.0 quarts and anticodon identity at 1 and 0 sites [8]. This places Phe^{11} entry into the code before Tyr^{11}. The restriction of same-family amino acids, including these latecomers, to codons with the same 5'-base, is credited to synthesome specificity for related tRNA when catalyzing closely related reaction sequences (§6).

Selection forces targeting residue polarity [4-6] were not required to explain the polar-residue cluster and they are not required for the non-polar cluster. Pre-LCA tRNA sequence identity [8] and 5'-base invariance rule [7] were noted to place Phe^{11} in the NUN cluster, independent of its hydrophobicity (ΔF_{T}, -3.7 kcal/mol) [23]. With $UU^{U}{}_{C}$ assigned to Phe^{11}, it shares a 4-set with Leu^{8}, another non-polar amino acid ($\Delta F_{T} = -2.8$ kcal/mol). Leucine^{8} path-distance places its entry into the code before Phe^{11}. Pre-LCA tRNA identity and pathway evidence show Leu^{8} captured a Ser^{4} tRNA, cognate with UUN, in a cofactor exchange (§3.2). Capture of CUN or AGN, assigned to Ser^{4} earlier (§3.1), likely involved prohibitive fitness costs. Apparently, UUN, and possibly CUN, initially coded for Ser^{4}, a nonhydrophobic amino acid. α–Amino acid Met^7/Ile^7 pathway intermediates also evidently acquired AUN, as a stage-2 START signal (§6.1). They include aspartate-semialdehyde^3 and homoserine^4 with respective ΔF_T values of 0.773 and 0.386 kcal/mol; estimated from $\Delta F_T = 2.587 + 2.818$ Z, where Z = - log P, P being a hydrophobicity parameter [48]. Indicative of the direction of protein evolution [1, 2], hydrophobicity increased during Met^7 and Ile^7 pathway extension; $\Delta F_T(Met^7) = -3.4$ kcal/mol, $\Delta F_T(Ile^7) = -3.1$ kcal/mol.

The non-polar NUN cluster is centered on 7-8 step stage-2 amino acids, Met^7, Ile^7, Leu^8, flanked by Val^5 and post-expansion amino acid, Phe^{11} (Fig. 4). Path-distances indicate it formed in a narrow interval of code evolution. This contrasts with the polar-residue cluster, which formed by 1-2 and 10-13 step residues. NAN and NUN cluster path-distance distributions differ significantly, $\chi_1^2 = 4.22$, $p = 4.0x10^{-2}$. As path-distances increased from 1 to 8 steps increasingly non-polar residues entered the code [1, 2]: mean ΔF_T (kcal/mol) = 6.58 (Asp^1, Glu^1, Asn^2 Gln^2), 0.43 (Ala^2, Hse^4, Pro^4, Ser^4), –1.86 (Gly^5, Cys^5, Val^5), and –3.10 (Met^7, Ile^7, Leu^8). Proteins plainly became more hydrophobic during expansion from the small NH_4^+ Fixers Code, with triplets recruited in a NAN → NUN direction [1]. Proteins that partition with the cell membrane had emerged by the final stage of code expansion [20].

Path-distance evidence demonstrates that polar and non-polar residue clusters formed at different stages of code evolution. They consequently conserve the imprint of protein evolution from solely polar residues to non-polar residue sequences. Path-distance evidence also shows NAN and NUN triplets, respectively, initiated and terminated early code growth (stages-1and –2). The path-distance principle [1, 2] thus clarifies why codons for polar and non-polar residues respectively cluster at NAN and NUN sets. Selection targeting residues with mixed polarity, in principle, could form clusters at any of 48 non-

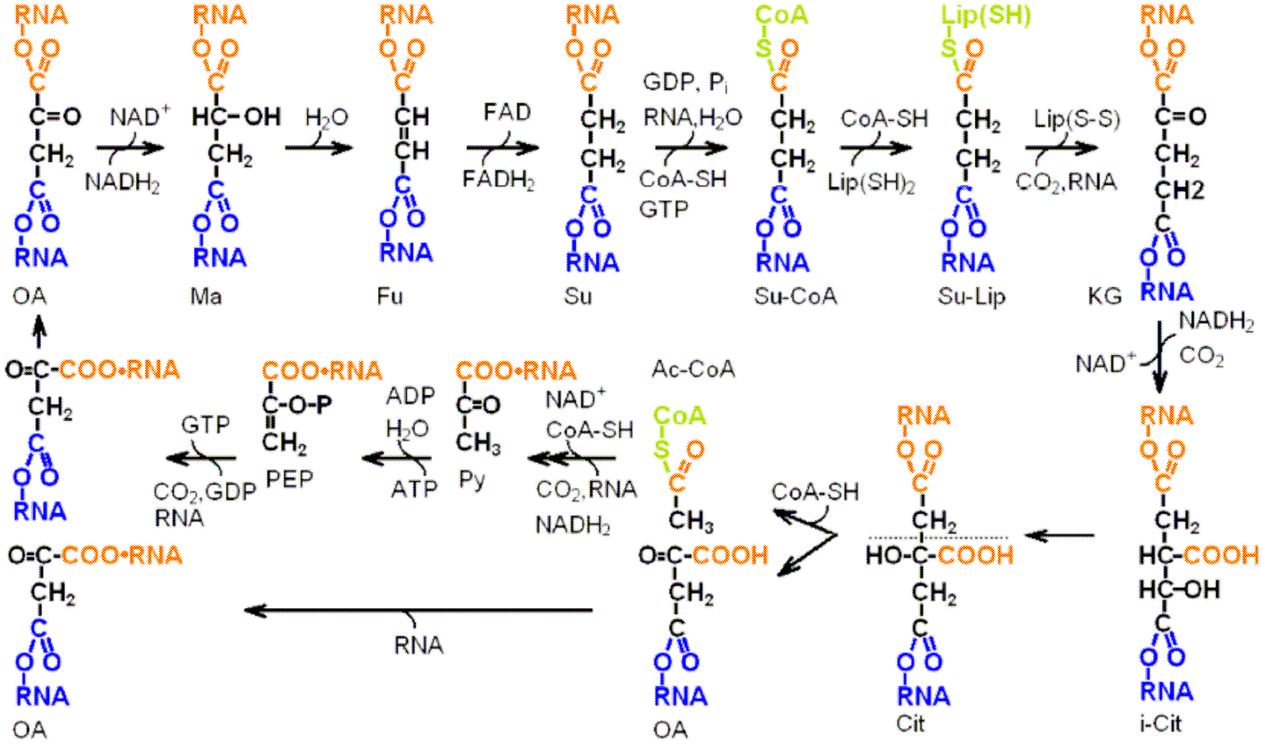


Figure 6. Distribution of potential RNA cofactor attachment sites among citrate cycle components. A free terminal carboxyl at C3 of oxaloacetate (OA) is a cycle invariant, consistent with attachment to an RNA cofactor or scaffold. A second free terminal carboxyl group, at C2 of OA, is quasi-invariant; coenzyme A (CoA) reacts with it in succinate (Su) and in citrate (Cit) to form succinyl- (Su-CoA) and acetyl-CoA (Ac-CoA), respectively. Ma, malate; Fu, fumarate; Su-Lip, succinyl-lipoate; KG, 2-ketoglutarate; i-Cit, isocitrate; Py, pyruvate; PEP, phosphoenolpyruvate. With RNA as a scaffold, each turn of the cycle converts 4 CO2 molecules to OA in a replicated OA-RNA strand.

overlapping triplet sets within the code (Table S1, feature 17).

Intermediates in amino acid biosynthesis, with few exceptions, contain a free α-carboxyl, consistent with masking by a pre-LCA tRNA cofactor [8]. Since these pathways mainly originate in the CC [1, 2], it is significant that OA C3-carboxyl is invariant and only succinyl-CoA, succinyl-lipoate, and acetyl-CoA displace its C2-carboxyl (Fig. 6). Citrate cycle components were evidently also attached to a pre-LCA RNA cofactor or scaffold. From pre-LCA tRNA cofactor identities, central metabolism segments were notably incorporated into amino acid synthesis pathways, upstream of Pyr, 3PGA, and PEP (Fig. 3).

Integrating metabolic pathways and replication [49] becomes conceivable. Each CC rotation would convert $4CO_2$ to an OA in a new OA-RNA strand (Fig. 6). Back-tracking from the D-ribose-5-phosphate scaffold in RNA leads to cyclization of D-ribulose-1,5-diphosphate in the reductive pentose cycle (RPC). The RPC has an invariant phosphate, consistent with an attached scaffold, and it autocatalytically produces a 3-PGA every three rotations. A double-strand 'ladder' of linear polyphosphate chains crosslinked by 2-carboxy-3-oxo-ribulose molecules (triose pairs), at the replication site, constitutes a replicative-form of polytriose-phosphate (PTP). With the PTP replicator located in the RPC, it is an apparent antecedent of RNA [50]. Phosphorylated formose cycle components, glyceraldehydes-3-phosphate and dihydroxyacetone, in the RPC link PTP replication to this spontaneously autocatalytic cycle, fueled by a 1-C molecule (formaldehyde). With no invariant sites, the formose cycle provides a pre-replication source for replicating metabolic pathways.

References

1. Davis BK. 1999 Evolution of the genetic code. *Prog. Biophys. Mol. Biol*. **72**, 157-243. (doi: 10.1016/S0079-6107(99)00006-1)
2. Davis BK. 2007 Making sense of the genetic code with the path-distance model. In *Leading-edge Messenger RNA Research Communications* (ed. MH. Ostrovisky) pp. 1-32. New York: Nova Science
3. Woese CR. 1965 Order in the genetic code. *Proc. Natl. Acad. Sci. USA.* **54**, 71-75. (doi: 10.1073/pnas.54.1.71)
4. Sonneborn TM. 1965 Degeneracy of the genetic code: extent, nature, and genetic implications. In *Evolving Genes and Proteins* (eds. V. Bryson, HJ. Vogel) pp. 379-397. New York: Academic Press.
5. Freeland SJ, Knight RD, Landweber LF, Hurst LD. 1998 Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511-518. (doi: 10.1093/oxfordjournals.molbev.a026331)
6. Ardell D, Sella G. 2001 On the evolution of redundancy in genetic codes. J. Mol. Evol. **53**, 269-281. (doi: 10.1007/s002390010217)
7. Taylor FJR, Coates D. 1989 The code within codes. *Biosystems* **22**,177-187. (doi: 10.1016/0303-2647(89)90059-2)
8. Davis BK. 2008 Imprint of early tRNA diversification on the genetic code: Domains of contiguous codons read by related adaptors for sibling amino acids. In *Messenger RNA Research Perspectives* (ed. T. Takayama) pp. 35-79. New York: Nova Science.
9. Saks ME, Sampson JR. 1995 Evolution of tRNA recognition systems and tRNA gene

sequences. *J. Mol.Evol.* **40**, 509-518. (doi: 10.1007/BF00166619)

10. Davis, BK. 2011 Genetic code domains conserve the imprint of tRNA cofactors encoded to specify cognate amino acid synthesis. (url: http://www.archive.org/details/GeneticCodeDomains)

11. Bailly M, Giannouli S, Blaise M, Stathopolous C, Kern D, Becker D. 2006 A single tRNA base pair mediates bacterial tRNA-dependent biosynthesis of asparagine. *Nuc. Acids Res.* **34**, 6083-6094. (doi: 10.1093/nar/gkl622)

12. Cork JM, Purugganan MD. 2004 The evolution of molecular genetic pathways and networks. *BioEssays* **26**, 479-484. (doi: 10.1002/bies.20026)

13. Palioura S. 2011 *RNA-dependent selenocysteine biosynthesis in eukaryotes and Archaea*. Ann Arbor: ProQuest UMI

14. Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A, von Haeseler A. 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**, 673-679. (doi: 10.1126/science.2497522)

15. Michal, G. 1992 *Biochemical Pathways*. 3rd Edition, Penzberg: Boehringer Mannheim

16. Diaz-Lazcoz Y, Henaut A, Vigier P, Risler JL. 1995 Differential codon usage for conserved amino acids: evidence that the serine codons TCN were primordial. *J. Mol. Biol.* **250**, 123-127. (doi: 10.1006/jmbi.1995.0363)

17. Doctor VM, Oro J 1972 Non-enzymatic decarboxylation of aspartic acid. *J. Mol. Evol.* **1**, 326-333. (doi: 10.1007/BF01653961)

18. Wong JT-F. 1975 A coevolution theory of the genetic code*. Proc. Natl. Acad. Sci USA.* **72**, 1909-1912. (doi: 10.1073/pnas.72.5.1909)

19. Davis BK. 2005. Coevolution theory of the genetic code: is the precursor-product

hypothesis invalid? *BioEssays.* **27**, 1308. (doi: 10.1002/bies.20332)

20. Giege R, Sissler M, Florentz C. 1998 Universal rules and idiosyncratic features in tRNA identity. *Nuc. AcidsRes.* **26**, 5017-5035. (doi: 10.1093/nar/26.22.5017)

21. Bretscher MS, Goodman HM, Menninger JR, Smith JD. 1965 Polypeptide chain termination using synthetic polynucleotides. *J. Mol. Biol.* **14**, 634-639. (doi: 10.1016/S0022-2836(65)80219-4)

22. Tolstrup N, Toftgard J, Engelbrecht J, Brunak S. 1994 Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. *J. Mol. Biol.* **243**, 816-820. (doi: 10.1006/jmbi.1994.1683)

23. Davis BK. 2002 Molecular evolution before the origin of species. *Prog. Biophys. Mol. Biol.* **79**, 77-133. (doi: 10.1016/S0079-6107(02)00012-3)

24. Jukes TH. 1973 Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophys. Res. Commun.* **53**, 709-714. (doi: 10.1016/0006-291X(73)90151-4)

25. Davis BK. 2009 On mapping the genetic code. *J.Theor. Biol.* **259**, 860-862. (doi: 10.1016/j.jtbi.2009.05.009)

26. Kyprides N, Overbeek R, Ouzounis C. 1999 Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* **49**, 413-423. (doi: 10.1007/PL00006564)

27. Lim V, Curran P. 2001 Analysis of codon:anticodon interactions within the ribosome provides new insights into code reading and genetic code structure. *RNA* **7**, 942-957. (doi:10.1017/S135583820100214X)

28 Dillon LS. 1973 The origins of the genetic code. *Botanical Rev.* **39**, 301-345. (doi: 10.1007/BF02859159)

29. Ronnenberg TA, Landweber LF, Freeland SJ. 2000 Testing a biosynthetic theory of the genetic code: Fact or artifact? *Proc. Natl. Acad. Sci USA* **97**, 13690-13695. (doi: 10.1073/pnas.250403097)

30. Hinegardner RT, Engelberger J. 1963 Rationale for a universal genetic code. *Science* **142**, 1083-1085. (doi: 10.1126/science.142.3595.1083)

31. Davis BK. 2004 Expansion of the genetic code in yeast: making life more complex. *BioEssays* **26**, 111-115. (doi: 10.1002/bies.10415)

32. Gutteridge A, Thornton JM. 2005 Understanding nature's catalytic toolkit. T*rends Biochem Sci.* **30**, 622-629. (doi: 10.1016/j.tibs.2005.09.006)

33. Munoz V, Serrano L. 1994 Intrinsic secondary structure propensities of the amino acids, using statistical φ-ψ matrices: Comparison with experimental scales. *Proteins* **20**, 301-311. (doi: 10.1002/prot.340200403)

34. Wilmot CM, Thornton JM. 1988 Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.* **203**, 221-232. (doi: 10.1016/0022-2836(88)90103-9)

35. Jurka J, Smith TF. 1987 β-Turns in early evolution: chirality, genetic code, and biosynthetic pathways. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 407-410. (doi: 10.1101/SQB.1987.052.01.047)

36 Walter KU, Vamvaca K, Hilvert D. 2005 An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280**, 37742-37746. (doi: 10.1074/jbc.M507210200)

37. Doudna JA, Lorsch JR. 2005 Ribozyme catalysis: not different , just worse. *Nature: Struc. Mol. Biol*. **12**, 395-402.(doi: 10.1038/nsmb932)

38. Suga H, Futai K, Jin K. 2011 Metal ion requirements in artifical ribozymes that catalyze aminoacylation and redox reactions. *Met. Ions Life Sci.* **9**, 277-297. (doi:

10.1039/9781849732512-00277)

39. De Duve C. 1988 The second genetic code. *Nature* **333**, 117-118 (doi: 10.1038/333117a0)

40. Blaise M, Bailly M, Frechin M, Behrens MA, Fischer F, Oliveira CLP, Becker HD, Pedersen JS, Thirup S, Kern D. 2010 Crystal structure of a transfer-ribonucleoprotein particle that promotes asparagine formation.*EMBO J* **29**, 3118-3129. (doi: 10.1038/emboj.2010.192)

41. Osawa, S. (1995) *Evolution of the genetic code,* Oxford: Oxford University Press.

42. Allwood A C, Walter MR, Burch IW, Kamber BS. 2007 3.43 billion-year-old stromatolite reef from the Pilbara Craton of Western Australia**:** Ecosystem-scale insights to early life on Earth. *Precambrian Res.* **158**, 198-227. (doi: 10.1016/j.precamres.2007.04.013)

43. McGuinness E. 2010 Some molecular moments of the Hadean and Archaean aeons: retrospective overview from the interfacing years of the second and third millennia. *Chem. Rev.* **110**, 5191-5215. (doi: 10.1021/cr050061l)

44. Crick FHC. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367-379. (doi: 10.1038/213119d0)

45. Trifonov EN. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139-151. (doi: 10.1016/S0378-1119(00)00476-5)

46. Miller S L, Orgel L. 1974 *The Origin of Life on the Earth* Engelwood-Cliffs: Prentice-Hall.

47. Dayhoff MO, Schwartz RM, Orcutt BC. 1978 A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (ed. MO Dayhoff), Vol. **5**, pp. 345-352 Silver Spring: National Medical Foundation.

48. Black SD, Mould DR. 1991 Development of hydrophobicity parameters to analyze

proteins which bear post- or cotranslational modifications. *Anal. Biochem.* **193**, 72-82. (doi: 10.1016/0003-2697(91)90045-U)

49. Orgel LE. 2008 The implausibility of metabolic cycles on the prebiotic Earth. *PLoS Biol* : **6(1)**, e18. (doi: 10.1371/journal.pbio.0060018)

50. Davis BK. 2012 Replicative-form of poly(triose-phosphate. (url: http://www.archive.org/details/Replicative-formOfPolytriose-phosphate)

Supplement

Brian K. Davis
Research Foundation of Southern California, Inc., La Jolla, California, USA
davis@resfdnsca.org

Table S1. Genetic code features with interpretations based on amino acid synthetic order in reconstructed tRNA-dependent pathways. aa, amino acid; superscript, amino acid path-distance; N, any standard base; X, Z coding bases; R, purine; Y, pyrimidine, •, degenerate 3'-base read by single tRNA bearing a U34; aaRS, aminoacyl-tRNA synthetase; LCA, Last Common Ancestor; PDP, path-distance principle; OA, oxaloacetate; Pyr, pyruvate; PGA, 3-phospho-glycerate; PEP, phosphoenolpyruvate.

code feature	interpretation
1. Codons assigned to aa sharing the same precursor tend to cluster in the same code region [1].	Code structure (domains) shows aa synthesis was initially tRNA-dependent. Pre-LCA tRNA phylogenetics revealed tRNA for same-family aa shared an ancestral tRNA [2]. Cofactor/adaptor tRNA for a new aa thus had anticodons/codons nearest-neighbor to those of aa sharing the same precursor and related pathways.
2. Probability experiments indicate aa were initially allotted triplets four-at-a-time. Later, new aa received a doublet, or single codon [3].	Earlycomer aa, synthesized on 2-8 step paths, have 7 of 8 codon 4-sets (3'-base degenerate) in code. All 6 latecomer aa (9-14 step paths) are encoded by a doublet, or single codon, shared with a short-path aa or stop signal [2, 4, 5] - Arg^9 4-set CGN reflects an α-aa intermediate. Path-distances confirm 4-sets preceded assignment of double and single codons to aa.
3. All 8 'codon 4-sets' have a G, or C, as 5'- and/or mid-base [6].	Pre-code translation on a poly(A) strand [4, 5] led to early tRNA species having a U34 (universal bp-forming anticodon 5'-base) and to early aa acquiring codon 4-sets (feature 2). Codon sets with a 3'-Y that lacked a G, or C, at 5'-, or mid-site, were ambiguous when read by a tRNA with a U34 [7]. Reducing the tRNA reading range from a codon 4-set to doublet (3'-Y, or 3'-R) [7], by modifying U34, restored coding fidelity. One doublet then became available for reassignment to an incoming aa, splitting the G/C deficient codon 4-set.
4. Genetic code antiquity implies aa originated by reductive organo-synthesis [3, 8]	C-atom oxidation no. decreases linearly with increases in path-distance among 14 early aa (1-8 step paths) [5]. in accord with initial reductive organo-synthesis. Appearance of a selectively permeable protein/lipid cell membrane [9] accounts for no further decrease among 6 latecomer aa (9 - 14 step paths).

Table S1 (continued).

code feature	interpretation
5. Triplets with a mid-A code for hydrophilic aa, while those with a mid-U code for the most hydrophobic aa in proteins [10].	NAN triplets code for 7 aa and a STOP signal. They include 4 first generation aa: 2 diacids and their amides with 1–2 step paths (Fig 4). NUN triplets encode 5 highly hydrophobic aa; 4 expansion phase aa (5-8 step paths) and post-expansion aa, Phe^{11}. Path-length evidence reveals the code evolved toward a hydrophobic attractor (membrane formation - feature 4), through a series of mid-base recruitments from NAN → NCN → NGN → NUN, preserving code compactness during expansion. This outcome also minimizes the risk of substitution between polar and non-polar residues in early proteins [11, 12]
6. Similar codons for same-family aa suggests aa synthesis involved tRNA cofactors during code formation [3, 13].	Subdivision of the code into domains containing specific combinations of same-family aa/phylogenetically related pre-LCA tRNA species/contiguous codons (Fig. 1b) conserves the imprint of extensive tRNA-dependent aa synthesis during code formation [2].
7. Codon mid-base has most coding capacity. 3'-Base has least [14].	Codon mid-site recruitment added 10 aa (2-8 step paths) and 12 4-sets, during expansion from the NH_4^+ Fixers Code. Whereas, 5'- and 3'-site recruitment, respectively, added only 4 aa (1-2 step paths) and four 4-sets initially, and 6 aa (9-14 step paths) and reassiged 5 doublets and 1 single [4, 5].
8. Codons for same-family aa exhibit 5'-base invariance [15].	5'-Base invariance among codons for same-family aa reflects codon mid-base recruitment during code expansion beyond the small NH_4^+ Fixers Code (Figs. 1b, 4) with NAN triplets as codons. Anticodon mid-base substitutions during diversification of cofactor/adaptor tRNA species [2] thus expanded the code from AAN and CAN triplets for Asn^2 and Gln^2, respectively. to ANN and CNN code for Asn^2, Thr^6, Met^7, Ile^7, Arg^9, and Lys^{10}, from the Asp^1 family, and Gln^2, Pro^4, His^{13}, and Arg^9 (a chimera combining both families), from the Glu^1 family.
9. Each of the six smallest aa in proteins were assigned a codon 4-sets [15].	GCN, ACN, CCN, UCN, GGN, and GUN encode Ala^2, Thr^6, Pro^4, Ser^4, Gly^5, and Val^5 with mean mol. wt. of 103 and path-distance of 4.3 steps. Earlycomer aa acquisition of stable (low error) 4-sets, read by U34-bearing (universal-pairing wobble site base) tRNA [4, 5] fits with a 'first in, best encoded' rule during code expansion.

Table S1 (continued).

code feature	interpretation
10. tRNA species with complementary anticodons show elevated base complementarity at acceptor stem site, N2 [16].	tRNA N2 is a G, C rich site. Base complementarity at N2 thus occurs when tRNA species differ at this site. tRNA with complementary anticodons are heterogeneous, because they generally come from different code domains [17]. tRNA from the same domain are related and read contiguous (usually non-complementary) codons (Fig. 1b).
11. Nucleotide-like aa have long synthesis paths [18].	Formation of His^{13} and Trp^{14} on long pathways indicates that the protein takeover of ribozymal reactions occurred late in code formation. tRNA-dependent aa synthesis likewise persisted until advanced stages of code formation [2].
12. Subdivision of codon 4-sets provided an error-suppression strategy [7].	Subdivided 4-sets GAN, AAN, CAN, UAN, AGN, UGN, AUN, and UUN are prone to mid-base ambiguity in translation, when a Y:Y bp is the wobble-site pair [7]. The PDP places a U34 in early tRNA (Fig. 4b), so this accounts for the pattern of subdivided 4-sets in the code. Initial reading ambiguity conceivably mitigated the risk of mutation to an unassigned triplet, which can block translation [19].
13. Residues that gain frequency with protein phylogenetic depth were earlycomers to the code [20].	Gainer-aa Ala^{2}, Val^{5}, Gly^{5}, Ile^{7}, Asp^{1}, Ser^{4}, and Asn^{2} have 1-7 step paths, making them code earlycomers; His^{13} is the sole exception [5]. Four o f 6 loser-aa, Lys^{10}, Phe^{11}, Tyr^{11}, and Trp^{14} have 9-14 step path, making them code latecomers. Thus, the proposed residue disequilibrium [20] and PDP [2, 4, 5] show broad agreement.
14. Asp^{1} and Asn^{2} gain frequency with phylogenetic depth, while their homologues Glu^{1} and Gln^{2} lose it [20].	Post-divergence compensation for under-representation of Glu^{1}/Gln^{2} in pre-LCA proteins reasonably accounts for their negative frequency-shift with phylogenetic depth. Pre-LCA tRNA phylogenetics show tRNA^{Gln} contributed tRNA for synthesis of only 2 aa, with Glu^{1} as precursor. In constrast, Asp^{1} produced 14 coded aa with tRNA cofactor/adaptors derived from tRNA^{Asn}.
15. Changes in aa residue frequency with phylogenetic depth indicate GNN triplets encoded aa before ANN [21].	Consistent with depth results, GNN encode aa, Asp^{1}, Glu^{1}, Ala^{2}, Gly^{5}, Val^{5}, with a shorter mean path-distance than ANN encoded aa, Asn^{2}, Ser^{4}, Thr^{6}, Ile^{7}, Met^{7}, Arg^{9}, Lys^{10}: 2.8 steps versus 6.4 steps. NAN encoded NH_4^+ fixer/N-donor aa, Asp^{1}, Glu^{1}, Asn^{2}, Gln^{2}, with a mean path-distance of 1.5 steps notably predate the GNN aa, in accord with NH_4^+ fixers being the first generation of aa [4, 5, 17].

Table S1 (continued)

code feature	interpretation
16. Five of 6 α-aa in the 4.6 billion yr. Murchison meteorite have codons in the GNN set [22].	The Murchison meteorite showed abiogenic sources of α-aa existed before Earth formed. It contained 5 aa (Asp^1, Glu^1, Ala^2, Gly^5, Val^5) with GNN codons and mean path-distance of 2.8 steps, placing them among the earliest aa to enter the code (feature 15). Code structure excludes abiogenic sources of α-aa from a role in the origin of proteins, however. Each code domain conserves a specific combination of same-family aa/related tRNA/contiguous codons (Figs. 1b, 4). Thus, they retain evidence of an extensive network of tRNA-dependent aa synthesis pathways during code formation [2].
17. Code clusters of polar and non-polar aa with NAN and NUN triplets, respectively, result from error-minimizing, pro-polarity-homology selection forces [23, 24].	The code contains 24 non-overlapping triplet sets comparable to NAN/NUN: 4 x 3 (5'-base) + 4 x 3 (mid-base). Each pair can have either a polar/non-polar or non-polar/polar aa arrangement. Thus, 48 sets of code clusters could optimize residue polarity-homology. In contrast, the PDP accounts for polar and non-polar aa, respectively, having NAN and NUN codons. Four of 7 aa with NAN codons have 1-2 step paths (Asp^1, Glu^1, Asn^2, Gln^2), placing them in the first code [1]. The remaining 3 aa (Lys^{10}, Tyr^{11}, His^{13}) form on 10-13 step paths, making them post-expansion additions to the code. NUN triplets code fot 5 non-polar aa (Val^5, Ile^7, Met^7, Leu^8, Phe^{11}) with 5-11 step paths, with most entering the code in its late expansion stage. Addition of increasingly hydrophobic aa during code expansion (NAN ↠ NUN) accompanied appearance of membrane proteins [5, 9]. As significantly different aa path-distance distributions characterize polar and non-polar clusters, the code conserves evidence of major shifts in the direction of protein evolution during code formation.
18. Charged aa have codons with a purine mid-base base and hydrophobic aa codons have a pyrimidine mid-base [25].	Proteins contain 5 charged residues. NAN triplets specify 4 (Asp$^-$, Glu$^-$, Lys$^+$, His$^{+/o}$) placing them in the hydrophilic aa cluster. Asp$^-$ and Glu$^-$ helped seed formation of this cluster (feature 17). Codons for each of these 4 aa are within their code domain or quasi-domain. Arg$^+$ has codons CGN and AGR, consistent with having both Glu$^-$ and Asp$^-$ as a precursor. Acquisition of CGN is credited to end-product transfer, on displacing ornithine$^+${5, 26]. The location of Arg codons fits with synthetic order among Glu-derived aa and direction of code expansion [5]. Among codons with a pyrimidine mid-base, highly hydrophobic aa, Val, Ile, Met, Leu, Phe, cluster on NUN. NUN triplets were reassigned last, as the code evolved toward a hydrophobic attractor: formation of membrane proteins [4, 5].

Table S1 (continued)

code feature	interpretation
19. Clusters of physically homologous aa [26] and codon-like ligand-binding bases in anti-aa aptamers [27] suggest direct aa recognition by codons.	Differences in aa path-distance reveal clusters of hydrophilic and hydrophobic aa originated at different stages of code evolution (Fig. 4b). The direction of protein evolution thus accounts for these clusters; invoking direct aa recognition by codons is thus unjustified. NMR spectra of an anti-Arg aptamer binding site [28] show no significant elevation in Arg codon (binomial) frequency. Single nucleotide frequencies of other aa in anti-aa aptamers also show no significant clustering of aa-specific triplets [29].
20. Codon 3'-base displays most coding degeneracy [14, 30].	With 64 triplets in the standard code, there are 43 more than required for 20 aa and a stop signal. 3'-Base degeneracy in eight 4-sets, 1 triple, and 13 doublets contributes 91 per cent (39/43) of the surplus. Code regions displaying 3'-base degeneracy pre-date post-expansion code formation, when the codon 3'-site encoded long-path (9-14 step) basic and aromatic aa [5].
21. Prokaryotes widely utilize a tRNA cofactor for Asn^2 and Gln^2 synthesis, but never for Asp^1 and Glu^1 [31-33].	Prokaryote reliance on tRNA-dependent synthesis of Asn^2 and Gln^2 (Fig. 1a) fits with both pathways being the protected root [34] of a once extensive network of pre-LCA tRNA-dependent aa pathways. As precursors, Asp^1 and Glu^1 synthesis was not tRNA-dependent [2, 35]. Code domains retain the imprint of this network, and half aa in the code still derive from Asp^1 or Glu^1 [4].
22. The origin of two classes of synthetase enzymes, each specific for 10 aa, whose tRNA binding and A76 acylation sites differ, is possibly linked to the origin of the code [36].	Phylogenetic and substrate distribution evidence indicate that Class I and II aaRS evolved by radiation from Glu-RS-I and Asp-RS-II, respectively [2]. Code domains and pre-LCA tRNA phylogenetics indicate Glu^1 and Asp^1 were precursors in an extensive network of tRNA-dependent aa synthesis pathways during code formation. This links dual aaRS classes to aa precursor (Asp^1, Glu^1) duality [2, 35].
23. The size of the aa alphabet corresponds to the chromatic number of the genetic code [37].	With 20 aa and 48 distinguishable codons (allowing for 3'-Y degeneracy), the code could form an ideal color map [37]. Each neighboring codon (more than one 'point' of similarity) then code for a different aa (color). All base changes thus produce an aa substitution, making the 'ideal' code error-prone. Codon degeneracy in the standard code arises in 8 4-sets, 1 triple, and 13 doublets, reducing transcription and translation errors.

Table S1 (continued)

code feature	interpretation
24. Steps 8 - 11 in Ile^7 synthesis duplicate reactions in Val^4 synthesis [38]	Splicing the whole Val path onto the Thr^6 path to form Ile^7, implies a cassette of ribozymes preceded the present suite of enzymes: acetolactate synthase, acetohydroxy acid isomero-reductase, dihydroxy acid dehydratase, and aminotransferase. Recruitment of the Val^4 cassette by a tRNA^{Ile} cofactor could then co-opt the Val-pathway in a single step.
25. The standard genetic code is virtually universal [30]	Code invariance in the long post-divergence interval reflects the elevated risk of any change being lethal, as genome size increased [39, 40]. An exponential fall-off in the number of codons assigned per reaction step, in aa pathways extending over 4-steps, reveals the tempo of code evolution declined gradually before the code 'froze' [5].
Code Features Revealed by the Path-Distance Principle	
26. Triplet coding sites were recruited in a 5'- → 3'-direction (Fig. 4), producing three distinct phases in code formation [4, 5, 35].	*5'-site*: 16 XAN triplets initially coded for 4 NH_4^+ fixer (1-2 step) aa and a STOP signal. The coding limit for a 1-site code (4-letter nucleotide alphabet) is not exceeded, given GA• (ambiguously) coded for both Asp$^-$ and Glu$^-$. *mid-site*: Its recruitment added 10 mainly small alkyl-chain (2-8 step) aa. With 15 aa , including Met (START) and Arg intermediates, and a STOP signal, the expansion code specified a 16 -letter alphabet – consistent with 2-site limit (4 x 4). *3'-site*: Its recruitment added 6 large basic and aromatic (9-14 step) aa. With 20 aa (dual purpose START) and STOP signal, the 3-site code specified a 21 letter alphabet - far smaller than its coding limit (4 x 4 x 4). A lack of triplets to reassign, due to risk of lethal code mutations linked to genome size, are credited with halting code growth (features 7 25).
27. 5' → 3' recruitment of codon sites during code formation (feature 26), parallels the direction of template translation.	Crick's continuity principle [30], broadly interpreted from the perspective of the PDP, requires that pre-code translation randomly assembled poly(Asp, Asn, Glu, Gln) from N- → C-terminus, while proceeding in a 5' → 3' direction on a poly(A) template [4, 5]. Recruiting codon sites in a 5' → 3' direction thus paralleled the direction of early translation.

Table S1 (continued)

code feature	interpretation
28. An invariant mid-A and degenerate 3'-base occupy non-coding sites in the first codons (Fig. 5).	Limiting the first code to triplets with mid-A confined it to a compact set of 16 codons. They were readable with as few as 4 tRNA species, bearing a U34 (universal bp forming wobble-site base) and U35. Mutation to one of 48 (¾ total triplets) unassigned triplets, which block translation [19], resulted solely from a mid-base (⅓ sites) substitution in the first code [4].
29. The first code helps define pre-code translation [4].	Back-tracking from A-rich codons, type-ID tRNA, and NH_4^+ fixer/donor aa of the first code indicates proteins originated within a primal NH_4^+ fixing/distribution system [5]. A single ancestral adaptor, tRNA-ID$^{(Asp/Glu/Asn/Gln)}_{UUU}$, cognate with AAA triplets, in a poly(A) strand, evidently produced random sequence amide-bearing, polyanionic oligopeptides of aa homologues - Asp$^-$, Asn, Glu$^-$, Gln.
30. Acidic aa residues have short paths, while basic aa have long paths [4]	Asp$^-$, Glu$^-$ form on 1-step paths, placing these diacid aa in the first generation of coded aa (Fig. 4). Basic aa, Arg$^+$ and Lys$^+$, with 9 - 10 step paths, represent post-expansion, latecomers. Early entry of acidic aa into the code and exclusion of basic aa, mirrors the ubiquity of multianionic reactants in central metabolism. An early reliance on multianionic proteins likely ceased, when porous protocells gave way to selectively permeable cells [9].
31. Seven of 8 codon 4-sets code for earlycomer aa (2 – 8 step paths). In contrast, all 6 late-comer aa (9-14 step paths) share a 4-set with an early-comer aa or STOP signal - Arg^9 is sole exception (Fig. 4).	Earlycomer aa Ala^2, Thr^6, Pro^4, Ser^4, Gly^5, Val^5, Leu^8 have codon 4-sets GCN, ACN, CCN, UCN, GGN, GUN, CUN. In contrast, doublets AAR, CAY, UAY, AGR, UUY code for post-expansion aa Lys^{10}, His^{13}, Tyr^{11}, Arg^9, Phe^{11}, and Trp^{14} has single codon, UGG; Arg^9 has sole 4-set, CGN, and it is credited to path-extension [41]. Each shares a 4-set with an earlycomer aa, or STOP signal. The pattern of intact and subdivided codon 4-sets acquired by short- and long-path aa corroborates that aa synthetic order was a determinant of time-order of code entry.

Table S1 (continued)

code feature	interpretation
32. Codon mid-base correlates with earlycomer aa path-distance [4, 5].	Mid-base triplet sets NAN, NCN, NGN, and NUN respectively code for earlycomer aa sets (Asp^1, Glu^1, Asn^2, Gln^2), (Ala^2, Thr^6, Pro^4, Ser^4), (Gly^5, Ser^4, Cys^5), and (Val^5, Ile^7, Met^7, Leu^8), with mean path-distances of 1.5 → 4.0 → 4.7 → 7.0 steps (Fig. 4). Recruiting triplets through successive mid-base substitutions preserved code compactness during its expansion (feature 7).
33. Pre-divergence proteins preferentially conserve short-path aa [9].	Conserved sites in pre-LCA proteins have aa from an earlier stage of code formation than aa at variable sites [9]. Ancient proteins thus conserve evidence that aa synthetic order determined the time-order of aa addition to the code.
34. Ferredoxin antecedent, Pro-Fd-5, has a structure suggesting some early proteins acted as surface adaptors [9].	Pro-Fd-5 is a 23-residue ferredoxin antecedent with a stage-5.6 residue profile (mid-expansion code). It has a negatively charged 7-aa N-terminus 'foot' linked to a [4Fe-4S] electron transfer center [42]. Pro-Fd-5 could thus anchor its cofactor to a cationic mineral surface within a protocell, prior to cells equipped with a selectively permeable plasma membrane, comprising a protein/phospholipid fluid mosaic.
35. aa synthesis intermediates conserve the attachment site for a tRNA cofactor [2].	All aa intermediates, except for His^{13} and Trp^{14} (at steps 12, 13), bear a free α-carboxyl (Fig. 3). This group is masked by cofactor in tRNA-dependent aa synthesis. Its ubiquity among aa intermediates, together with code structure and pre-LCA tRNA phylogenetics provide compelling evidence that aa synthesis was tRNA-dependent during code formation [2].
36. A dicarboxylated intermediate is a feature of Leu^8 and Arg^9 synthesis (Fig. 3c,d).	Isopropyl-malate and arginine-succinate are dicarboxylated intermediates in Leu^8 and Arg^9 synthesis, respectively. Each occurs at a putative transition point from tRNA-IA to tRNA-II, cognate with UUN codons, in Leu synthesis, and from tRNA-ID to tRNA-IA in Arg synthesis. These observations corroborate former evidence of tRNA-dependent amino acids synthesis during code formation [2].

code feature	interpretation
37. Ser^4 and Leu^8 precursors differ, but both charge type-II tRNA cognate with neighboring codons (Fig 1).	Consistent with Leu^8 acquiring a Ser^4 tRNA at step-5 of synthesis (Fig. 3c), tRNA-$II^{Ser}_{3'AGU}$ is ancestral to tRNA-$II^{Leu}_{3'AAU}$ based on a pre-LCA sequence identity of 5.7 quarts [2]. This reconciles the difference in aa synthesis family (Fig. 1), with the sharing of type II tRNA cognate with neighboring codons.
38. Diacid homologues Glu^1 and Asp^1 are each precursor to an aa family (Fig. 4) and each has a different class of synthetase [43].	Pre-LCA sequence identity indicates that tRNA species arose from tRNA-IA^{Asn}, or tRNA-ID^{Gln} [2, 35]. An initial reliance on tRNA-dependent aa synthesis pathways implies Asp^1 and Glu^1 were the primal precursors of all aa incorporated into proteins. Precursor and cofactor duality provide a source of synthetase duality, in contemporary acylation and, inferentially, in ribozymal synthetases at initiation of tRNA-dependent aa synthesis.
39. aa substrate distribution among class I and II synthetases (aaRS) is centered around Glu^1 and Asp^1, respectively [2].	Path-distances of motif and signature-segment residues place the origin of aaRS class I and II late in code formation [4]. A mean aaRS-I substrate aa molecular weight of 150 (range, 117-204) approximates Glu^1 at 146. An aaRS-II aa substrate mean of 123 (75-165) is also near 132 of Asp^1. Thus, aa substrate distributions and aaRS phylogenetics fit with radiation from Glu-RS-I and Asp-RS-II. Each was likely preceded by a ribozymal synthetase specific for one of the precursor aa [2].
40. tRNA for aa not derived from Asp can exhibit close identity to $tRNA^{Asn}$ [35].	Pre-LCA tRNA-$IA^{Ala}_{3'CGU}$ identity with tRNA-IA^{Asn}_{UUU} is 11.0 quarts [2]. Evidently, Pyr-family aa (Ala^2, Val^5, Leu^8) synthesis once included conversion of Asp-RNA^{Ala} → Pyr-$tRNAA^{Ala}$ via deamination and decarboxylation (Fig. 2a). Pre-LCA $tRNA^{Phe}$ identity with $tRNA^{Met}$, and $tRNA^{Ser}$ with $tRNA^{Asn}$ implies an OA-tRNA → Pyr-tRNA → PEP-tRNA → PGA-tRNA path existed.
41. Asp^1 is precursor of 6 aa in biosynthesis, and Glu^1 to 3 3 aa. Pre-LCA tRNA phylogenetics shows Asp^1 to be precursor of 15 aa during code formation [2,35].	Pre-LCA tRNA identities reveal Asp was initially precursor to 15 aa (Fig.3). In this respect, pre-LCA identity for $tRNA^{product\ aa}$ vs. $tRNA^{Asp}_{3''CUG}$ is [2]: 1.48 ± 0.33 quarts (m ± s.e.m.), n = 15, $p = 4^{-1.48} = 0.128$ NS, and $tRNA^{product\ aa}$ vs. $tRNA^{Asn}_{3''UUG}$ 7.94 ± 1.10 quarts, n = 14, $p = 1.67\text{x}10^{-5}$ S ($p = 2.0\text{x}10^{-6}$ for $tRNA^{'Asn'}_{UUU}$), where Asp product aa were drawn from Fig. 3.

Table S1 (continued)

code feature	interpretation
42. Over-representation of Asp derived aa in proteins (features 14, 41) breaks pair symmetry of aa homologues in NH_4^+ Fixers Code	This asymmetry points to functional specialization. Glu was preferred NH_4^+ donor, by more than 200-fold, in amidation of Asp-tRNA-IA^{Asn} and Glu-tRNA-ID^{Gln} [44], consistent with Asp being preferred precursor and tRNA-IA^{Asn} preferred cofactor. The IA core group in tRNA^{Asn} also breaks symmetry in the first code, facilitating Asn cofactor/adaptor recognition during code expansion.
43. Codon 4-set GA^R_Y encodes both precursor aa, Asp^1 and Glu^1, in the most ancient part of the code (Fig. 1b).	GA^R_Y and adjoining codon sets NA• and GN• , are the most ancient parts of the early code (aa $\ngtr$ 8 steps). Their mean aa path-distances (steps) are: GA^R_Y, 1 (2 aa), NA• , 1.5 (4 aa), and GN• , 2.8 steps (5 aa). By contrast, the 9 non-adjoining codon sets to GA^R_Y encode 7 aa, with mean path of 5.9 steps (Fig. 3). GA^R_Y aa, Asp$^-$ and Glu$^-$, notably fix and distribute NH_4^+ and have catalytic potential, consistent with a pre-translation role.
44. α-Amine addition occurs near the end of long aa synthesis pathways (Fig. 4).	Intermediates of long-path, ring-bearing aa, Phe^{11}, Tyr^{11}, Trp^{14}, typically contain an α-carboxyl, but lack an α-amine, preventing mis-incorporation of an intermediate. Long-path basic aa (Arg^9, Lys^{10}) also avoid mis-incorporation this way, until the last 2-3 steps.
45. Met^7 initiates translation, but is a late expansion phase aa (Fig. 5).	All intermediates of Met^7, and sibling Thr^6, are α-amino acids. Thus, translation conceivably started at a 5'-AU• codon, as early as the first code (Fig. 5). This opened the way for assembly of proteins with a defined residue sequence and expansion beyond the small, locally-phased NH_4^+ Fixers Code [2].
46. Code domains conserve evidence of an expanded role for tRNA during code formation [2].	Identity elements in tRNA acceptor-, D-, and T-arms recruit a synthetase and transamidase in tRNA-dependent amide aa synthesis [45]. This links tRNA with aa 'selection' at the origin of translation [2]. In addition to aa path-recognition, pre-LCA tRNA functioned as an adaptor in early translation. Its dual roles likely accounts for tRNA being larger than initially envisioned [46].
47. Thermostable and thermolabile aa formed the first generation of coded aa (Fig. 5).	Asp^1 and Glu^1 are thermostable, similar to other abiogenic molecules [47]. Both were present at the origin of proteins as free aa. In contrast, Asn^2 and Gln^2, and most intermediates, were bound to a tRNA cofactor (features 6, 50). Unlike the abiogenic scenario [48], lactam formation by amide aa thus posed no risk. Lability of the ester bond joining an aa-, or intermediate-, to tRNA would have constrained physical conditions at the origin of translation [5, 49].

Table S1 (continued)

code feature	interpretation
48. NH_4^+ fixer/donor aa, Asp^1, Glu^1, Gln^2 have type-ID tRNA, while Asn^2 has a type-IA tRNA [4],	Backtracking from the NH_4^+ fixer/donor code leads to pre-code translation with a type-ID 'universal' acceptor/cofactor, tRNA-ID$^{Asp,Glu,Asn,Gln}_{UUU}$, cognate with AAA triplets in poly(A), to assemble random poly(Asp,Asn,Glu,Gln) oligopeptides [2]. As tRNA-IA$^{Asn}_{UUU}$ also had a UUU anticodon, acquisition of a IA core group distinguished it from ancestral tRNA, in competing for AAA codons.
49. Same-domain tRNA retain the same core group throughout aa pathway extension (Fig. 4a).	tRNA for same -family aa, with path-distances spanning all stages of code evolution, retain the same core group. This supports descent from a common ancestral tRNA, in each code domain [2, 5]. Core group invariance also implies secondary structure of tRNA, with shared pathways, was constrained during code formation [35].
50. The conserved imprint of pre-divergence tRNA specific for same-family aa reveal they were structurally related and cognate with similar codons (Fig. 1c).	Analysis of pre-LCA tRNA sequences, with post-divergence variations filtered-out, established that tRNA, specific for aa synthesized from the same precursor, diversified from a common ancestral tRNA [2]. The kinship between pre-LCA tRNA species for synthetically related aa is obscured, when post-divergence tRNA variations - ⅔ of total [50]– are not excluded [51].
51. Code structure conserves the imprint of coordinated tRNA/aa pathway/code evolution (Fig. 4).	Code domains combine related pre-LCA tRNA species, contiguous codons, and same-family L-aa [2]. Partitioning the code into domains shows the growth of nascent aa synthesis pathways [52] and code formation were coordinated with the diversification of cofactor/adaptor tRNA molecules. Code structure excludes racemic mixtures of abiogenic aa [47, 53] from forming the first proteins.
52. Successive recruitment of codon 5', mid, and 3' site accompanied addition of generically different aa to the code, in each phase of its evolution (Fig. 4b).	Episodic shifts in the aa alphabet reveal changes in the forces driving protein evolution and codon recruitment. aa encoded by the 5'-site of XAN triplets, Asp$^-$, Asn, Glu$^-$, Gln, formed proteins assuredly functioning at the point of N atom entry into an RNA-based metabolism. Mid-base recruitment led to mainly small, alkyl-chain aa, Ala, Pro, Gly, Val, Ile, Leu, facilitating synthesis of proteins that partition with a membrane lipid layer. Appearance of cells with selectively permeable membranes is linked to codon 3'-site recruitment and incorporation of large basic and cyclic chain aa, Arg$^+$, Lys$^+$, Phe, Tyr, His, Trp [9]. Assignment of all available triplets at each stage, motivated by error-minimization, contributed to the stepwise nature of code evolution.

Table S1 references

1. Nirenberg M, Caskey T, Marshall R, Brimbacombe R, Kellogg D, Doctor B, Hatfield D, Levin J, Rottman F, Petska S. et al. 1966 The RNA code and protein synthesis. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 11-24. (doi: 10.1101/SQB.1966.031.01.008)
2. Davis BK. 2008 *Imprint of early tRNA diversification on the genetic code: Domains of contiguous codons read by related adaptors for sibling amino acids.* In Messenger RNA Research Perspectives (ed. T. Takayama) pp. 35-79. New York: Nova Science.
3. Dillon LS. 1973 The origins of the genetic code. *Botanical Rev*. **39**, 301-345. (doi: 10.1007/BF02859159)
4. Davis BK. 1999 Evolution of the genetic code. *Prog. Biophys. Mol. Biol.* **72**, 157-243. (doi: 10.1016/S0079-6107(99)00006-1).
5. Davis BK. 2007 *Making sense of the genetic code with the path-distance model.* Leading-edge Messenger RNA Research Communications (ed. MH. Ostrovisky), pp. 1-32. New York:Nova Science.
6. Dunnill P. 1966 Triplet-nucleotide-amino-acid pairing: a stereochemical basis for the division between protein and non-protein amino acids. *Nature* **210**, 1267-1268. (doi: 10.1038/2101267a0)
7. Lim V, Curran P. 2001 Analysis of codon:anticodon interactions within the ribosome provides new insights into code reading and genetic code structure. *RNA* **7**, 942-957. (doi: 10.1017/S135583820100214X)
8. Wachterhauser G. 1992 Groundworks for an evolutionary biochemistry: the iron- sulphur world. *Prog. Biophys. Mol. Biol.* **58**, 85-201. (doi: 10.1016/0079-6107(92)90022-X)
9. Davis BK. 2002 Molecular evolution before the origin of species. *Prog. Biophys. Mol. Biol.* **79**, 77-133. (doi: 10.1016/S0079-6107(02)00012-3)

10. Woese CR. 1965 Order in the genetic code. *Proc. Natl. Acad. Sci. USA*. **54**, 71-75. (doi: 10.1073/pnas.54.1.71)

11. Sonneborn TM. 1965 *Degeneracy of the genetic code: extent, nature, and genetic implications.* In Evolving Genes and Proteins (eds. V. Bryson, HJ. Vogel) pp. 379-397. New York: Academic Press.

12. Freeland S.J, Hurst L. 1998 Load minimization of the genetic code: history does not explain the pattern. *Proc. R. Soc. Lond. B* **265**, 2111-2119. (doi: 10.1098/rspb.1998.0547)

13. Wong JT-F. 1975 A coevolution theory of the genetic code*. Proc. Natl. Acad. Sci USA.* **72**, 1909-1912. (doi: 10.1073/pnas.72.5.1909)

14. Perlwitz MD, Burks C, Waterman MS. 1988 Pattern analysis of the genetic code. *Advan. App Math.* **9**, 7-21. (doi: 10.1016/0196-8858(88)90003-6)

15. Taylor FJR., Coates D. 1989 The code within codes. *Biosystems* **22**, 177-187. (doi: 10.1016/0303-2647(89)90059-2)

16. Rodin AS, Szathmary E, Rodin SN. 2009 One ancestor for two codes viewed from the perspective of two complementary modes of tRNA aminoacylation. *Biology Direct* **4-4**: 1-30. (doi: 0.1186/1745-6150-4-4)

17. Davis, BK. 2011 Genetic code domains conserve the imprint of tRNA cofactors encoded to specify cognate amino acid synthesis. (url: http://www.archive.org/details/GeneticCodeDomains)

18. Garrett, RH., Grisham, CM. 1999 *Biochemistry* San Diego: Saunders. (doi:)

19. Bretscher, MS., Goodman, HM., Menninger, JR., Smith JD. 1965. Polypeptide chain termination using synthetic polynucleotides. *J. Mol. Biol.* **14**, 634-639. (doi: 10.1016/S0022-2836(65)80219-4)

20. Brooks DJ, Fresco JR, Lesk AM, Singh M. 2002 Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code.

Mol. Biol. Evol. **19**, 1645-1655. (doi: 10.1093/oxfordjournals.molbev.a003988)

21. Brooks DJ, Fresco JR. 2003 Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. *Gene* **303**, 177-185. (doi: 10.1016/S0378-1119(02)01176-9)

22. Kvenvolden K, Lawless J, Pering K, Peterson E, Flores J, Ponnamperuma C, Kaplan IA, Moore C. 1970 Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. Nature **228**, 923-92610. (doi: 10.1038/228923a0)

23. Freeland SJ, Knight RD, Landweber LF, Hurst LD. 1998 Early fixation of an optimal genetic code. *Mol. Biol Evol.* **17**, 511-518. (doi: 10.1093/oxfordjournals.molbev.a026331)

24. Ardell D, Sella G. 2001 On the evolution of redundancy in genetic codes. J. Mol. Evol. **53**, 269-281. (doi: 10.1007/s002390010217)

25. Biro JC, Benyo B, Sansom C, Szlavecz A, Fordos G, Micsik T, Benyo Z. 2003 A common periodic table of codons and amino acids. *Biochem. Biophys. Res. Comm*. **306**, 408-415. (doi: 10.1016/S0006-291X(03)00974-4)

26. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC. 1966 On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 723-731. (doi: 10.1101/SQB.1966.031.01.093)

27. Yarus M. 2000 RNA-ligand chemistry: A testable source for the genetic code. RNA 6, 475-484. (doi: 10.1017/S1355838200002569)

28. Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M. 1996 Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy..*Science* **272**,1343-1347. (doi: 10.1126/science.272.5266.1343)

29. Davis BK. 2008. Comments on the search for the source of the genetic code. *Messenger RNA Research Perspectives* (ed. T. Takeyama) pp. 1-8. New York: Nova Science,.

30. Crick FHC. 1966 Genetic code – yesterday, today, and tomorrow. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 3-5. (doi: 10.1101/SQB.1966.031.01.007)

31. Wilcox M, Nirenberg M. 1968 Transfer RNA as a cofactor coupling amino acid synthesis with that of protein. *Proc. Natl. Acad. Sci. USA* **61**, 229-236. (doi: 10.1073/pnas.61.1.229)

32. Danchin A. 1989 Homeotopic transformation and the origin of translation. *Prog. Biophys. Mol. Biol.* **54**, 81-86. (doi: 10.1016/0079-6107(89)90010-2)

33. Blaise M, Bailly M, Frenchin M, Behrens MA. Fischer F, Oliviera LP, Becker HD, Pedersen JS, Thirup S, Kern D. 2010 Crystal structure of a transfer-ribonucleoprotein particle that promotes asparagine formation. *EMBO J*. **29**, 3118-3129. (doi: 10.1038/emboj.2010.192)

34. Cork JM, Purugganan MD. 2004 The evolution of molecular genetic pathways and networks. *BioEssays* **26**, 479-484. (doi: 10.1002/bies.20026)

35. Davis BK. 2009 On mapping the genetic code. J.Theor. Biol**.** 259, 860-862. (doi: 10.1016/j.jtbi.2009.05.009)

36. Williams TA, Wolfe KH, Fares MA. 2009 No Rosetta stone for a sense-antisense origin of aminoacyl tRNA synthetase classes. *Mol. Biol. Evol.* **26**, 445-450. (doi: 10.1093/molbev/msn267)

37. Tlusty T. 2010 A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. *Phys. Life Rev*. **7**, 362-376. (doi: 10.1016/j.plrev.2010.06.002)

38. Rodwell VW. 1969 *Biosynthesis of amino acids and related compounds.* In Metabolic Pathways (ed. DM. Greenberg) **Vol. 3**, pp. 317-373, New York: Academic Press.

39. Hinegardner RT, Engelberger J. 1963 Rationale for a universal genetic code. *Science* **142**, 1083-1085. (doi: 10.1126/science.142.3595.1083)

40. Davis BK. 2004 Expansion of the genetic code in yeast: making life more complex.

BioEssays **26**, 111-115. (doi: 10.1002/bies.10415)

41. Jukes TH. 1973 Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophys. Res. Commun.* **53**, 709-714. (doi: 10.1016/0006-291X(73)90151-4)

42. Norgaard, H., Helt, SS, O, BL, Hagen, WR, Christensen, HEM. 2009 Spectroscopic characterization of evolutionary old ferredoxins. *J. Biol. Inorganic Chem.* ***14,*** Supp. 1.

43. Eriani, G., Delarue, M., Poch, O., Gangloff, J., Moras, D. 1990 Partition of aminoacyl-tRNA synthetases into two classes on the basis of two mutually exclusive sets of sequence motifs. Nature **347**, 203-206. (doi: 10.1038/347203a0)

44. Sheppard, K. 2007 *RNA-dependent Biosynthesis of Glutamine in Bacteria and Archaea*. Thesis. New Haven: Yale University.

45. Bailly M, Giannouli S, Blaise M, Stathopolous C, Kern D, Becker D. 2006 A single tRNA base pair mediates bacterial tRNA-dependent biosynthesis of asparagine. *Nuc. Acids Res.* **34**, 6083-6094. (doi: 10.1093/nar/gkl622)

46. Crick FHC. 1958 A note to the tRNA tie club. In M. B. Hoagland, 1960. *The Nucleic Acids*, (ed. N. Davison) **Vol. 3**, p. 349. New York: Academic Press.

47. Miller SL, Orgel L. 1974 *The Origin of Life on the Earth* Engelwood-Cliffs, New Jersey: Prentice-Hall.

48. Weber AL, Miller SL. 1981 Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol*. **17**, 273-284. (doi: 10.1007/BF01795749)

49. Davis BK. 1971 Chain propagation and polypeptide polymerization rate. *J. Theor. Biol*. **30**, 203-210. (doi: 10.1016/0022-5193(71)90045-2)

50. Eigen M, Lindemann BF, Tietz M, Winkler-Oswatitsch R, Dress A, von Haeseler A. 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**, 673- 679. (doi: 10.1126/science.2497522)

51. Sun F-J, Caetano-Anolles G. 2008 Evolutionary patterns in the sequence and structure of transfer RNA: a window into early translation and the genetic code. *PLoS One* **3**: e2799. (doi:)

52. Kyprides N, Overbeek R, Ouzounis C. 1999 Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* **49**, 413-423. (doi: 10.1007/PL00006564)

53. Trifonov EN. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139-151. (doi: 10.1016/S0378-1119(00)00476-5)

Table S2. Amino acid potentials for catalysis and protein structural features at indicated stages in genetic code formation. Potentials are the logarithm of the χ^2 probability, p, (corrected for continuity) for the distribution of a given residue between specified (hits) and non-specified (misses) protein sites versus the distribution for all other residues. The number of residues in these distributions were aggregates from 191 enzymes examined [1] to determine catalysis potentials, 279 proteins with α-helix and β-sheet structures [2], and 59 proteins with β-turns [3]. Residues whose frequency exceeded expectation were given a pro-potential of $-\log_{10} p$ and for a frequency under expectation an anti-potential of $\log_{10} p$ was assigned. Bold numbers denote potentials with $p < 10^{-2}$, $0 < p < 1$.

catalysis

code stage	amino acid	no. hits	no. misses	total	χ^2	$p(\chi^2)$	potential
1	Asp	165	6,963	7,128	196.09	1.49E-44	**43.83**
	Glu	122	7,633	7,755	52.64	4.00E-13	**12.40**
	Asn	43	5,399	5,442	0.11	0.74	-0.13
	Gln	21	4,380	4,401	6.75	0.01	**-2.03**
2	Ala	12	11,224	11,236	78.74	7.09E-19	**-18.15**
	Pro	9	6,222	6,231	37.20	1.07E-09	**-8.97**
	Ser	43	6,510	6,553	2.56	0.11	-0.96
	Val	8	8,745	8,753	62.36	2.86E-15	**-14.54**
	Cys	43	1,497	1,540	69.12	9.26E-17	**16.03**
	Gly	36	9,901	9,937	28.93	7.52E-08	**-7.12**
	Thr	50	6,986	7,036	1.32	0.25	-0.60
	Ile	2	6,792	6,794	55.64	8.71E-14	**-13.06**
	Met	2	3,112	3,114	55.64….	2.57E-06	**-5.59**
	Leu	7	10,727	10,734	55.64….	5.93E-20	**-19.23**
3	Arg	125	5,905	6,030	114.42	1.05E-26	**25.98**
	Lys	87	6,411	6,498	19.95	7.94E-06	**5.10**
	Phe	16	4,795	4,811	14.82	1.19E-04	**-3.93**
	Tyr	62	4,211	4,273	19.15	1.21E-05	**4.92**
	His	170	2,809	2,979	863.16	1E-189	**189.00**
	Trp	10	1,836	1,846	1.65	0.20	-0.70
	Total	1,033	122,058	123,098			

Table S2 (continued)

α−helix

code stage	amino acid	no. hits	no. misses	total	χ^2	$p(\chi^2)$	potential
1	Asp	732	2,795	3,527	6.56	0.01	-1.98
	Glu	1,128	2,396	3,524	193.05	6.85E-44	**43.16**
	Asn	437	2,308	2,745	71.37	2.96E-17	**-16.53**
	Gln	573	1,529	2,102	27.85	1.31E-07	**6.88**
2	Ala	1,774	3,248	5,022	515.42	4.19E-114	**113.38**
	Pro	435	2,361	2,796	81.00	2.26E-19	**-18.65**
	Ser	727	3,005	3,732	20.84	4.99E-06	**-5.30**
	Val	841	3,260	4,101	10.10	1.52E-03	**-2.82**
	Cys	158	866	1,024	29.57	5.39E-08	**-7.27**
	Gly	488	4,362	4,850	469.64	3.83E-104	**-103.42**
	Thr	609	2,994	3,603	68.93	1.02E-16	**-15.99**
	Ile	719	2,435	3,154	0.14	0.71	0.15
	Met	356	876	1,232	29.00	7.25E-08	**7.14**
	Leu	1,397	3,472	4,869	115.66	5.63E-27	**26.25**
3	Arg	731	1,876	2,607	47.39	5.81E-12	**11.24**
	Lys	920	2,534	3,454	35.46	2.60E-09	**8.58**
	Phe	472	1,948	2,420	12.93	3.23E-04	**-3.49**
	Tyr	396	1,780	2,176	24.12	9.03E-07	**-6.04**
	His	224	1,058	1,282	18.80	1.45E-05	**-4.84**
	Trp	193	704	897	0.46	0.50	-0.30
	Total	13,310	45,807	59,117			

Table S2. (continued)

β–turn (i)

code	**amino**	**no. hits**	**no. misses**	**total**	**χ^2**	***p*(χ^2)**	**potential**
stage	**acid**						
1	Asp	37	532	569	30.69	3.03E-08	**7.52**
	Glu	7	423	430	1.68	0.20	-0.71
	Asn	39	416	455	58.73	1.81E-14	**13.74**
	Gln	8	316	324	0.02	0.90	-0.05
2	Ala	11	800	811	5.83	0.02	-1.80
	Pro	13	421	434	0.03	0.86	0.07
	Ser	35	828	863	5.64	1.76E-02	1.75
	Val	4	649	653	11.09	8.70E-04	**-3.06**
	Cys	13	234	247	5.12	2.36E-02	1.63
	Gly	20	821	841	0.32	0.57	-0.24
	Thr	23	578	601	2.43	0.12	0.92
	Ile	2	376	378	6.39	0.01	-1.94
	Met	0	124	124	2.58	0.11	-0.96
	Leu	8	676	684	6.21	0.01	-1.90
3	Arg	1	257	258	4.64	0.03	-1.51
	Lys	9	586	595	3.12	0.08	-1.11
	Phe	6	328	334	0.82	0.37	-0.44
	Tyr	6	340	346	1.00	0.32	-0.50
	His	10	235	245	1.22	0.27	0.57
	Trp	4	145	149	0.04	0.83	-0.08
	Total	256	9,085	9,341			

Table S2 (continued)

β–turn (i + 1)

code stage	amino acid	no. hits	no. misses	total	χ^2	$p(\chi^2)$	potential
1	Asp	27	542	569	8.10	4.42E-03	**2.35**
	Glu	19	411	430	3.98	4.60E-02	1.34
	Asn	7	448	455	2.21	0.14	-0.86
	Gln	5	319	324	1.42	0.23	-0.63
2	Ala	23	788	811	5.03E-04	0.98	0.01
	Pro	29	405	434	24.53	7.30E-07	**6.14**
	Ser	46	817	863	22.31	2.32E-06	**5.63**
	Val	9	644	653	4.47	3.46E-02	-1.46
	Cys	5	242	247	0.27	0.61	-0.22
	Gly	11	830	841	6.69	9.68E-03	**-2.01**
	Thr	25	576	601	4.13	4.21E-02	1.37
	Ile	6	372	378	1.59	0.21	-0.68
	Met	0	124	124	2.60	0.11	-0.97
	Leu	12	672	684	2.40	0.12	-0.92
3	Arg	8	250	258	0.02	0.89	0.05
	Lys	16	579	595	2.91E-04	0.99	-0.01
	Phe	2	332	334	5.23	2.22E-02	-1.65
	Tyr	4	342	346	2.86	9.10E-02	-1.04
	His	2	243	245	2.84	9.19E-02	-1.04
	Trp	2	147	149	0.66	0.42	-0.38
	Total	258	9083	9341			

Table S2. (continued)

β–turn (I + 2)

code stage	**amino acid**	**no. hits**	**no. misses**	**total**	**χ^2**	***p* (χ^2)**	**potential**
1	Asp	41	528	569	42.80	6.06E-11	**10.22**
	Glu	17	413	430	1.94..	0.16	0.79
	Asn	23	432	455	8.49	3.58E-03	**2.45**
	Gln	7	317	324	0.25	0.62	-0.21
2	Ala	19	792	811	0.42	0.516	-0.29
	Pro	1	433	434	9.90	1.66E-03	**-2.78**
	Ser	38	825	863	8.88	2.89E-03	**2.54**
	Val	6	647	653	8.16	4.29E-03	**-2.37**
	Cys	9	238	247	0.44	0.51	-0.29
	Gly	16	825	841	2.20	0.14	-0.86
	Thr	17	584	601	6.58E-04	0.98	0.01
	Ile	2	376	378	6.47	0.01	-1.96
	Met	1	123	124	1.13	0.29	-0.54
	Leu	12	672	684	2.40	0.12	-0.92
3	Arg	13	245	258	4.29	0.04	1.42
	Lys	13	582	595	0.58	0.45	-0.35
	Phe	6	328	334	0.86	0.35	-0.45
	Tyr	7	339	346	0.47	0.49	-0.31
	His	5	240	245	0.25	0.62	0.21
	Trp	5	144	149	3.76E-02	0.85	0.07
	Total	258	9,083	9,341			

Table S2. (continued)

β–turn (i + 3)

code stage	**amino acid**	**no. hits**	**no. misses**	**total**	**χ^2**	**$p(\chi^2)$**	**potential**
1	Asp	17	552	569	3.03E-02	0.86	0.06
	Glu	11	419	430	1.98E-02	0.89	-0.05
	Asn	17	438	455	1.26	0.26	0.58
	Gln	8	316	324	3.17E-02	0.86	-0.07
2	Ala	13	798	811	4.11	0.04	-1.37
	Pro	0	434	434	11.97	5.39E-04	**-3.27**
	Ser	19	844	863	0.96	0.33	-0.49
	Val	10	643	653	3.58	0.06	-1.23
	Cys	11	236	247	2.02	0.16	-0.81
	Gly	60	781	841	62.90	2.17E-15	**14.66**
	Thr	18	583	601	.3.91E-2	0.84	0.07
	Ile	5	373	378	2.57	0.11	-0.96
	Met	8	116	124	4.95	0.03	1.58
	Leu	14	670	684	1.20	0.27	-0.56
3	Arg	3	255	258	2.00	0.16	-0.80
	Lys	7	588	595	5.45	0.02	-1.71
	Phe	14	320	334	2.03	0.15	0.81
	Tyr	10	336	346	1.89E-03	0.97	0.02
	His	4	241	245	.0.83	0.36	-0.44
	Trp	11	138	149	.10.17	1.43E-03	**2.85**
	Total	260	9,081	9,341			

Table S2. (continued)

β–sheet

code stage	amino acid	no. hits	no. misses	total	χ^2	$p(\chi^2)$	potential
1	Asp	612	2,915	3,527	153.27	3.35E-35	**-34.47**
	Glu	695	2,829	3,524	82.40	1.11E-19	**-18.95**
	Asn	586	2,159	2,745	35.62	2.39E-09	**-8.62**
	Gln	490	1,612	2,102	9.64	1.90E-03	**-2.72**
2	Ala	964	4,058	5,022	141.10	1.53E-32	**-31.82**
	Pro	34	2,762	2,796	949.30	1.88E-208	**-207.72**
	Ser	993	2,739	3,732	0.23	0.63	0.20
	Val	1939..	2,162	4,101	1,004.26	2.13E-220	**219.67**
	Cys	384	640	1,024	67.38	2.24E-16	**15.65**
	Gly	453	4,397	4,850	780.26	1.05E-171	**-170.98**
	Thr	1379..	2,224	3,603	285.23	5.44E-64	**63.26**
	Ile	1356..	1,798	3,154	480.74	1.48E-106	**105.83**
	Met	360	872	1,232	5.54	1.86E-02	1.73
	Leu	1336..	3,533	4,869	3.73	5.34E-02	1.27
3	Arg	694	1,913	2,607	0.16	0.69	0.16
	Lys	836	2,618	3,454	7.91	4.92E-03	**-2.31**
	Phe	902	1,518	2,420	157.40	4.19E-36	**35.38**
	Tyr	845	1,331	2,176	183.70	7.56E-42	**41.12**
	His	378	904	1,282	..6.86	8.80E-03	**2.06**
	Trp	289	608	897	16.38	5.18E-05	**4.29**
	Total	15,525	43,592	59,117			

Table S2 references

1. Gutteridge A, Thornton JM. 2005 Understanding nature's catalytic toolkit. T*rends Biochem Sci.* **30**, 622-629. (doi: 10.1016/j.tibs.2005.09.006)
2. Munoz V, Serrano L. 1994 Intrinsic secondary structure propensities of the amino acids, using statistical φ-ψ matrices: Comparison with experimental scales. *Proteins* **20**, 301-311. (doi: 10.1002/prot.340200403)
3. Wilmot CM, Thornton JM. 1988 Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.* **203**, 221-232. (doi: 10.1016/0022-2836(88)90103-9)